"The single biggest problem in
communication is the illusion
that it has taken place."

George Bernard Shaw

Each Christmas I buy the *Private Eye* annual (an English satirical magazine) only to be slightly disappointed when much of the humor falls flat, yet I can watch the TV current affairs quiz 'Have I Got News for You' featuring its editor and be reduced to tears of laughter. Being at a live recording of the show is even more powerful. Why is this? Why is the experience and effect so different? Is it just the sense of occasion when I go to a live show or is there something more to shared experience?

We appear to learn more from lectures delivered in person than reading the lecturer's book, or even watching the same lecture recorded on video. Studies show children who are read to by their parents do better than if they are left to follow along with a CD. Two groups of children were tested on two made-up words used in a story. The children read to by their parents had an 80% recall rate, while children who followed the CD only 17%. This is a big disparity. The simplest explanation is that the children who were read to pay more attention. Are there other effects?

IMAX

Most scientists believe communication between humans is classical: words spoken in proximity have no more power than had we carefully written out what we wanted to say. Body language and tone of voice are simply useful tools to aid the transmission of this information. I'm going to explore the ways in which human face-to-face communication might exceed this traditional classical model. Let us look first at the bandwidth of communication between people.

Bandwidth

Let me give you a mental picture for what I mean by bandwidth. Imagine I am sitting in a darkened theatre enjoying one of my favorite comedians at the Edinburgh Festival – the biggest arts festival in the world. I phone a friend who is also a fan and let them listen in. Perhaps I even use the camera and surreptitiously point the phone at the comedian. My live experience is digitized, compressed and transmitted over the mobile network to my friend. He gets the same experience but at much-reduced bandwidth.

My friend has a similar but qualitatively different experience to mine. He cannot hear the degrees of loud and soft I hear, nor the full range of high and low frequencies forming the timbre of the comedian's voice; no sense of the smell of old armchairs or the heat of the audience around me. He is spared the strange stickiness my shoes meet on the floor of the auditorium and the occasional slosh of beer that hits me from a slightly inebriated neighbor. For the person at the other end of the phone, their view is of a tiny two-dimensional screen about 4 by 3 inches square. Of course, they can enlarge the picture, but then the pixilation dominates and it looks like an impressionist picture viewed close up. He has nothing like the same intensity of experience. Loss of bandwidth is something we can study mathematically and the reduction is enormous.

Video and Audio

The image of the comedy show is digitized by the camera and microphone; the video at 384,000 bits per second and the audio at 64,000 bps. Mathematical compression will be applied and the video will shrink to 30,000 while the audio drops down to 4,700. After compression, the whole experience amounts to around 40,000 bits per second. To put it in some perspective, a DVD would be 11.5 million bits per second, nearly 300 times the bandwidth.

My in-person experience has much higher bandwidth than even a DVD. It may even have infinite bandwidth. Physicists argue whether space-time is quantized but, for now, we will look at what would be needed to reproduce the experience faithfully on modern digital recording equipment.

Digitization

When something is converted to digital form, it goes through a number of steps. First, some way must be found to chop the thing into small parts in space and time. Then each of these parts is sampled with a sensor to give an electrical signal and, finally, this signal is measured and turned into a number.

Old microphones used carbon granules. As the sound waves passed through them, the granules were shaken and made better contact with each other. Connecting a battery across the granules gave a varying voltage. Modern microphones use a variety of technologies. The preference of most recording artists today is the electret microphone. A coil moves inside a magnet generating a varying voltage which is translated into a voltage as before.

Next we use a fast running clock and measure the voltage on each tick giving us a sequence of numbers. We have created a near perfect record of the sound, and we can prove this by recreating the sound through a loudspeaker. This is what happens every time you listen to your iPod.

To digitize film, each frame must be split in space as well as time. On each tick of the clock, a process scans left to right and top to bottom to form a one-dimensional stream of numbers that records the image. The system cuts the picture up into little elements called pixels, standing for picture elements. Each small square has its average color measured for red, green, and blue content coded as a number.

Digitization techniques have become the dominant way electronics work in the home, and digitization circuits are now ubiquitous.

Reality

How big is reality? Setting aside for a moment the problem that it might be infinite, we need to reproduce all the elements that go to make it up.

A normal DVD has an image of 720 by 576 pixels with 16 bits of color depth and a frame rate of 25 frames per second. The eye, however, is considerably better than this and a DVD does not fool it. HD video is 1900 by 1000 pixels with 32 bits of color depth and 100 frames per second. This is a great deal better – if you enjoy watching sport or nature documentaries, the additional resolution is amazing. This still falls far short of reality. An IMAX theatre gives a wrap-around image of about 10,000 by 7,000 pixels and comes closer to the average resolution of the human eye, estimated at about 30,000 by 20,000 pixels. But the eye cheats. It concentrates the rods and cones in the central portion of the retina. Although IMAX achieves the average pixel density of your eye, it comes nowhere near the peak density which is nearly 10,000 times greater.

For a truly equivalent experience, we would need about 320 million pixels *per* eye at a frame rate of 120 frames per second, allowing us full stereo synthesis. At this speed and resolution, we are matching the visual acuity of the eye and should be able to fool it completely. But there is one more problem to overcome: The image is not interactive. Move your head in the real world and the image will change. The objects in the foreground will vary their position in relation to the background, so-called motion parallax. Try it now, move your head and you will see that the book, or screen you are reading moves in relation to the background. In a simple digitized 3D image this will not happen. You *will* have a 3D image but you will not have a *real* image, a light field.

To create a real image you need to view a hologram or use head-tracking technology. A hologram records the light waves given off by an object in multiple directions rather than just the intensity of the light striking the camera through a single focal point. When you shine a laser back through the hologram, it regenerates the light waves as they would have originally come from the object. That light can be viewed from different directions, giving the impression of three dimensions rather than a mere two-dimensional photograph. There is often a limitation in viewing angle because the original photographic plate must wrap all the way around an object to capture the full 3D light field, but the illusion is very convincing.

A more effective way to create a real experience – and one with no restriction on viewing angle – is to construct the image in a computer and track the movement of your head. The computer can create the two-dimensional images each eye would see if the scene were truly three-dimensional. Computer software resolves motion parallax and a host of other elements, but to do so the computer must understand a model of the world so it can calculate how the scene would look from a particular angle and in the appropriate lighting. Some recent games such as Activision's 'Call of Duty' do this, and the experience is compelling.

Hologram

There are still problems. The image is stereo but planar. All the light coming into your eye comes from the screen a meter or so away. In the real world objects need you to change the focal length of your eye to bring them into sharp focus depending on their proximity. Try looking at your hand as you move it towards and away from your face, too close and your eye can't pull focus any further and it will blur. This mismatch between focal depth and the apparent distance implied by motion parallax is one of the reasons you can get headaches watching 3D images. There is something not quite right about them and your eye has to learn a new behavior.

Audio Field

Our poor friend at the end of the phone is listening to a mere 4700 bits per second rendition of the comedian. A compact disk is 64,000, 16-bit samples per second in stereo, over a million bits per second. So the information content of a mobile phone call is very low. It is a miracle you can understand speech at all over such a narrow channel, but this is made possible by two factors. First, human speech uses a limited range of frequencies. All the information in our voices lies within about two octaves centered on middle C. And, second, you can perform some clever mathematics to generate speech from seed information. For a given speaker the vowel 'a' might be 20% middle C, 50% F and 25% A, with a few other things thrown in for good measure. We can transmit this information and ask the computer at the other end to re-synthesize it. This is what happens when you listen to someone on a modern phone. You do not hear their actual speech, you listen to a computer synthesizer make a near approximation.

CD is no longer the gold standard for sound. Professional audio has standardized on 24 bit recording which is probably far beyond the limit of the human ear. An audio soundtrack is doing a good job at 2 million bits per second.

Sitting perfectly still in the middle of a room, each ear will pick up a different signal if the source is not directly in front of us. The two ears on a human head face a little forward, and the hair on your head slightly absorbs sound. We can calculate the source of the sound by the slight difference in the times at which it strikes each ear, and the variation in frequency content. We can use these two pieces of information to determine the direction from which a sound is coming. It was useful for our ancestors to be able to tell where the saber-tooth tiger was hiding. We can gain more accurate information by turning our heads from side to side. The differences in frequency and timing should vary as we do so and we gain a little more data to perform the calculation. If we walk through the room we sample yet more of the soundscape and this can be used to pinpoint the exact location of the source. As we move, we expect the sounds we hear to change subtly according to the mental model we use for locating objects in the soundscape.

To give the illusion of a soundscape modern systems use multiple microphones to capture the sound, so it can be reproduced on multiple speakers. Ideally, we would record a hologram of the sound but it is possible to record on thirty or so microphones and mix the tracks down to 5 or more channels giving us the sound experience we now expect at a modern cinema.

What is the Bandwidth of Life?

We have not yet talked about the other senses; smell, vibration, temperature, balance, wind chill, and touch. In all, there are over 25 senses that must be stimulated accurately to simulate reality. Just think for a moment how much information must be replayed to reproduce the sensation of bungee jumping off a bridge in the jungle or taking off into space, or giving birth. To digitize life completely, we need to stimulate every relevant nerve ending in the human body in real-time – skin, ears, eyes, balance, pain centers, and so on.

At the low end, a 'perfect' IMAX production would require 360 degree stereoscopic projection and the generation of a full sound field. This would take 3 Gigabits per second for the audio field and 5,600 terabits per second for the video field. This could be substantially reduced if the person wears virtual reality glasses to track their head and eye movements, but then you are substituting resolution with computer power.

At the high end, a team at the US Department of Energy's Fermilab estimate reality needs one hundred trillion samples per inch for a 'simple' quantum representation. If we look at the many worlds view of quantum mechanics, each photon hitting our eye can't be fully described by a single number. The photon may be entangled with other realities we should keep track of. This causes our picture of reality to become wildly complex. Everything we might see and experience is in some way a combination of possibilities, and these possibilities all interact. Real life is *very* complex.

Symbolic Communication

Computers have no concept of an in-person meeting. They communicate using purely symbolic methods in binary numbers. These have the same meaning whether communicated over a short piece of wire or using a fiber optic cable half way around the world. Computers never have to communicate understanding to each other because they use programs and a program can be perfectly transmitted. Body language is, of course, completely alien to them!

We know there are non-computable things; functions, numbers, musical melodies, and mathematical puzzles. Why would there not also be a place for non-computation in communication? David Deutsch has suggested human creativity is used to guess the 'program' running in someone's mind, and evolved so we can learn skills. Instead, might face-to-face communication be important because it lets us impart knowledge in a non-symbolic manner?

Hyper-communication

As with hyper-computing, hyper-communication is controversial. We instinctively know human communication is very different to computer communication. Face-to-face communications have a qualitatively different feel to them. My question is this. Is there more to face-to-face communication between human beings than the simple exchange of symbolic information?

Let us propose an experiment. I erect a 3D screen with a hi-fi surround sound system in a university lecture hall and deliver a lecture to a camera in the adjacent hall. Half the students see the lecture directly, and half remotely. With modern screens, it might be possible to set up the experiment so well that is difficult to tell which hall I am actually in. Is the experience of the remote students the same as the ones sitting in direct proximity with me? Do mirror neurons fire more strongly and pick up more information when you see me in the flesh, or is the feeling that a lecture is better when you are 'physically there' an illusion? You are perhaps less likely to fall asleep in my lecture if you are physically there because you are afraid I might walk over and hit you! What possible non-classical effects could be in play when you see an event or communicate in person that might make the communication different? Here are two potential differences:

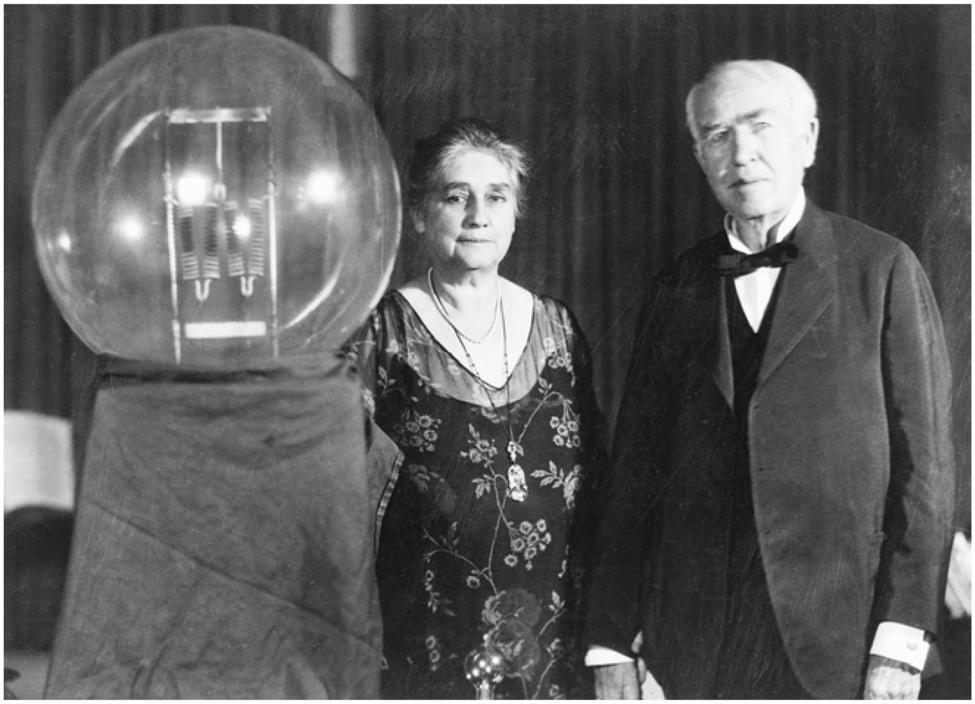
Information in a face-to-face encounter is continuous, not digitized. Continuous information is infinite in nature and does not have the finite limitation of digitized data. Of course, if we have digitized the sound at 24 bits and replayed it with extreme fidelity, there should not be any loss in information, but the interactivity of the soundscape is hard to simulate.

Light entering your eye contains information that could be quantum entangled with the object you are viewing. You become part of the system rather than merely an independent observer. It is difficult to see why this would produce a different quality of communication but it is testable. Set up the lecture experiment and measure the quality of understanding communicated between the parties.

If we believe our brains are super-Turing, then considering there might be some similar effects involved in human communication is not unreasonable, perhaps quantum effects play a role in communication. If we conclude our brains think classically, then we *probably* communicate classically.

Chapter 14

CREATIVITY



Thomas Edison, his wife and a Light Bulb

“Creativity is allowing yourself to make mistakes. Art is knowing which ones to keep.”

Scott Adams

“Invention is 1% inspiration and 99% perspiration.”

Thomas Edison

“Creativity is just connecting things. When you ask creative people how they did something, they feel a little guilty because they didn’t really do it, they just saw something. It seemed obvious to them after a while. That’s because they were able to connect experiences they’ve had and synthesize new things.”

Steve Jobs

The ancient Greeks believed there was no such thing as creativity. Our job, as humans, was to look at the earth and discover things about it. When we looked at light passing through water or built a boat to travel on it, we were discovering, not inventing. Shipwrights did not invent boats they were simply building inevitable forms. Everything there was to know already existed, we just hadn't realized it yet. Of course, Greek playwrights were busy 'creating' the first plays; tragedies, comedies and the like, but serious thinkers thought of them as documenting the human condition. It wasn't until the Renaissance, 1500 years later, that humans began to appreciate that they *create* knowledge, and this started us on our quest to understand creativity.

One of my childhood memories is sitting on the kitchen floor with a glass of water and surrounded by knives and milk bottles. I was trying to solve one of the problems from Edward de Bono's book on lateral thinking, *A Five-day Course in Thinking*. De Bono, now in his 80s, is a prolific writer with over 60 publications to his name – all aimed at making us more creative. His books pose a series of practical problems, each needing progressively greater creative intelligence. The particular problem I was trying to solve was to balance a glass of water on knives suspended from four milk bottles. It took me after 2 hours.

Steve Jobs shows the iPhone

Except for De Bono there is not much written about creativity in books or on the Internet, but if you dip into the video archive, the discussion really opens up. Perhaps this is a feature of creativity; it's easier to explain in person. Of course, I have taken on the writing task with this book but I have the benefit of modern day resources such as multimedia, interactivity, and the web.

Some people appear to have creativity in abundance and the things they create are truly wonderful. I'm thinking here of Picasso, Einstein, Mozart, Edison, or Maxwell, but a precise definition of creative thinking is hard to pin down. Here are some generally accepted categories:

Divergent Thinking

The first sort of creative thinking we recognize is divergent thinking, often called brainstorming. This is the art of coming up with ideas – lots of them. A quick way to test your skill is to take a minute, and list all the possible uses for a paperclip. Try it!

In 60 seconds write down all the uses for a paper clip you can think of. Time yourself.

◆ ◆ ◆

ANSWER WITHOUT READING ON

This is the classic test of creativity developed by J.P. Guilford in 1967. It is called the Alternative Uses Task. You can try the task with many objects: bricks, chairs, even water. How did you do on your first attempt? 8 to 10 uses is about average, 20 is extremely good. It's possible to teach most people to get near twenty and I'll show you how to do this in a moment.

Another test of idea generation is to draw 30 things in 30 circles. Thirty is such a large number it forces us to come up with some nutty ideas and break our natural tendency to self-censor.

For example, I'd like you to create logos or logo ideas, for a new coffee company in your circles. The test is best done without a time limit so now is the time to break off reading and make yourself a coffee. Then come back and draw 30 circles on a piece of paper. Fill in the circles.

MAKE A COFFEE, THEN START DRAWING.

◆ ◆ ◆

The aim of brainstorming is to remove our inhibitions and get us to generate a mass of ideas. In normal life, we tend to suppress ideas even before we are consciously aware of them. Sir Ken Robinson has researched creativity in children and found the ability to brainstorm reduces linearly with age. At five or six, children given one of these divergent thinking tasks come up with many creative solutions: fold the paper clip into a dinosaur, and use it to attack your friends, get two and use them as chop sticks. As adults, we tend to disqualify ideas. You could never fold a paperclip that tightly or accurately, we said, "a" paper clip not two. But, you *can* fold a paper clip tightly, and the room you are doing the test in has thirty paper clips and thirty people in it so just team up with a friend. I never said this was a solo task!

Do you see how you impose nonexistent rules on your thinking, particularly the implied rule of not working with others? I did not say this test was subject to examination conditions. The first twenty years of our lives teaches us to work alone on intellectual tasks, yet when we get to the workplace we can, and indeed must collaborate to succeed. Now you have an idea how to ace the paper clip test: don't censor yourself and don't imply rules I have not imposed!

TRY THE PAPER CLIP TEST AGAIN!

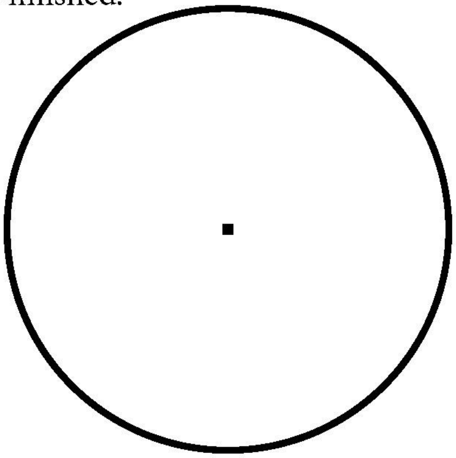
◆ ◆ ◆

Divergent thinking is rarely the final goal; it is rather a jumping off point for the creation of something new, like a solution to a mathematical puzzle, a painting, or a novel invention. The exercises allow us to explicitly see one of the early creative steps – idea generation before the pruning step. But most creative people often just create, they don't follow a scripted process. The term 'the creative process' is a great misnomer. There is no process that actually creates. Process merely puts us in the right frame of mind to do so. Processes are useful for framing a problem and ensuring we have all the right tools at our disposal: good crayons, some nice art paper, a hot cup of coffee. But process must be put to one side at the moment of actual creation.

Convergent Thinking

Convergent thinking is the opposite of divergent thinking. It focuses on discovering the final solution to a problem rather than generating precursor ideas. Some creative people only use this method, avoiding laborious processes such as brainstorming.

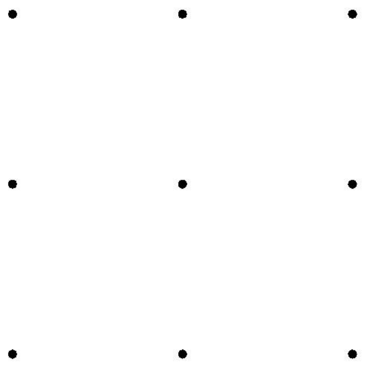
Tests of skill for convergent thinking generally pose puzzles where there is only one correct answer, but one that requires a non-linear step. Here's a really simple convergent thinking puzzle to try. It only requires a piece of paper and pen. Draw a circle on a piece of paper with a dot in the center. It should look like this. Do NOT take the tip of the pen off the paper until you are finished.



TRY ANSWERING IN YOUR OWN TIME

◆ ◆ ◆

Another famous, but clichéd, problem is of you to draw four straight lines through these nine points without lifting the pen from the paper. Can you do it?



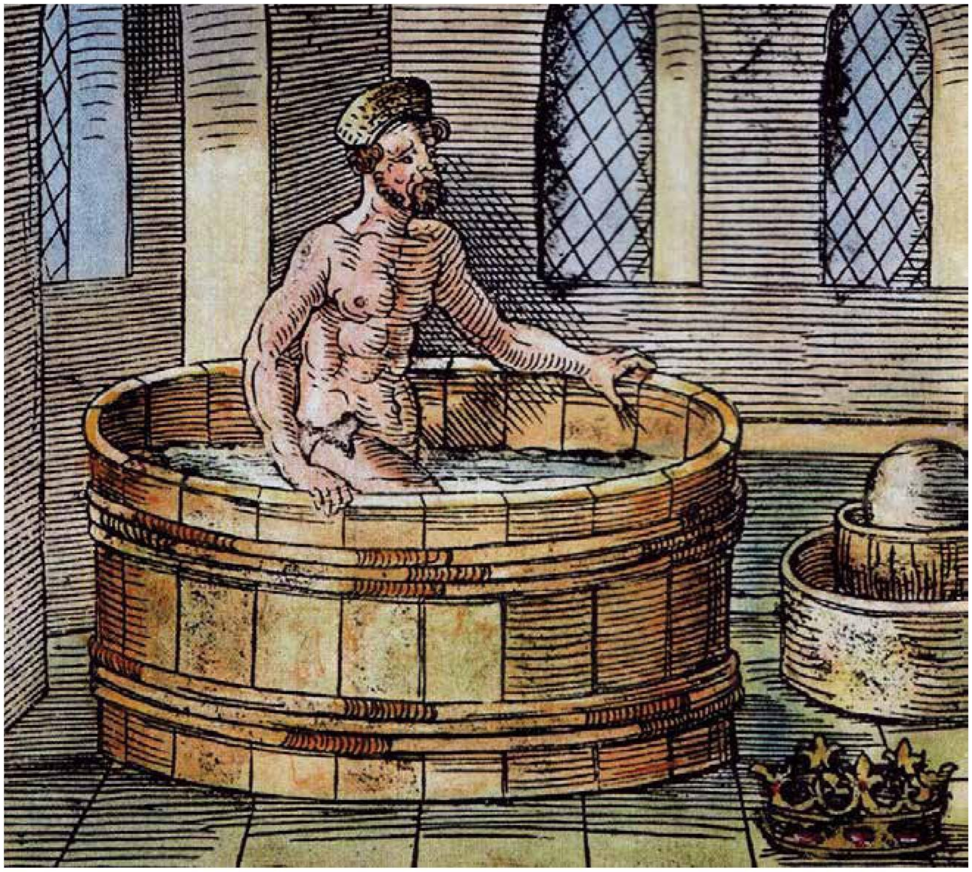
TRY ANSWERING IN YOUR OWN TIME

◆ ◆ ◆

I won't put the answers here, or even a hint. It's quite famous and many of you will be familiar with them. The answers are buried on the website, and for those of you who already know the solution, there are some alternative problems. If you can't immediately solve a problem, think about it overnight. It's worth seeing what your brain will do while you are asleep!

The Science of Creativity

The first person to theorize about the creative process was Graham Wallas, the co-founder of the London School of Economics. In his book *The Art of Thought,* he proposed a five-step model for creative thinking. First preparation, when you become fully acquainted with the problem and its domain. Then incubation; walk the dog or make a cup of tea. After the meditative incubation phase you may get a gut feeling that a solution is on its way. Wallas called this third step intimation. It's left out of many modern versions of his theory, but I think it's an important step. Shortly



Eureka

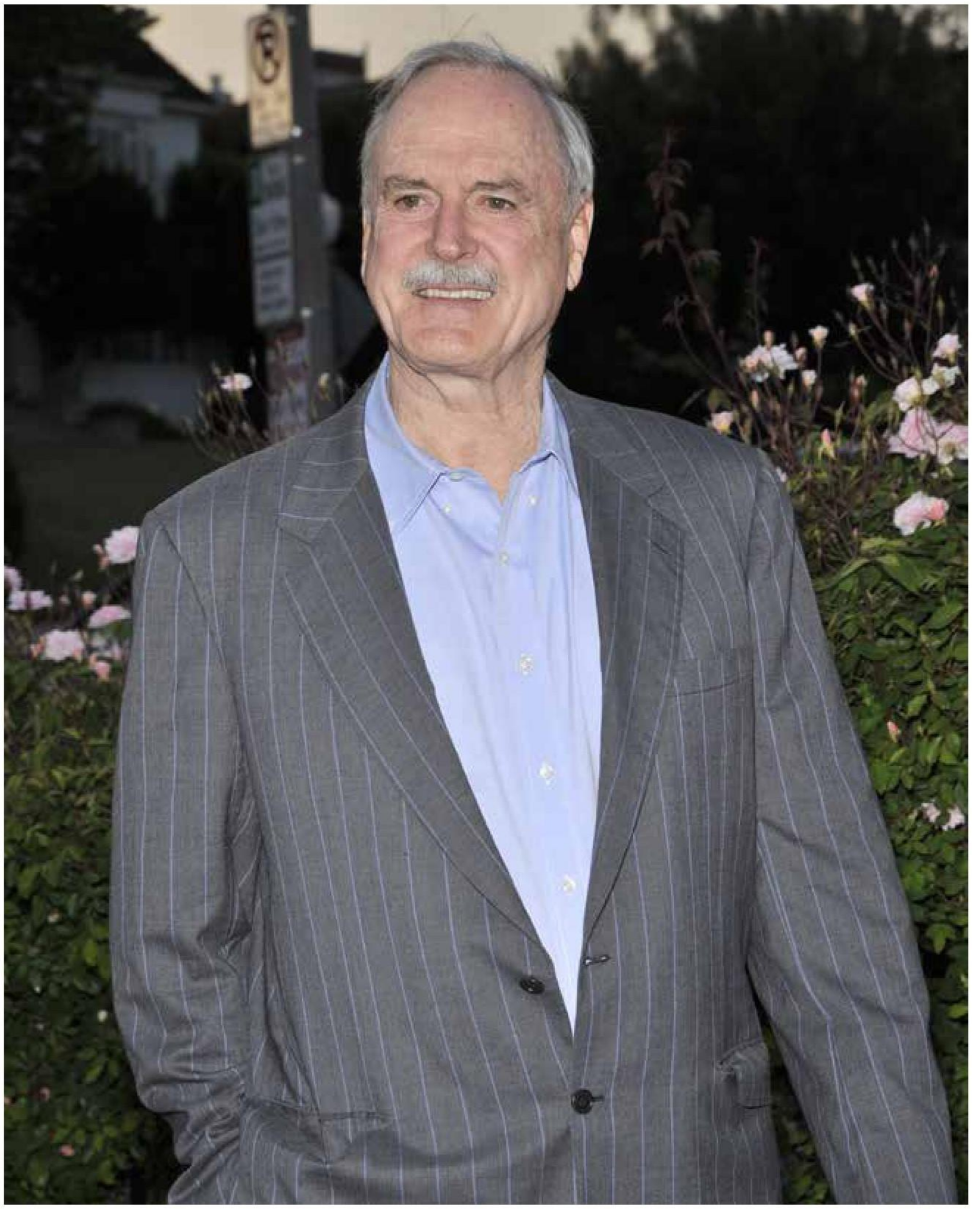
after this you get that Eureka moment – illumination or insight where the creative idea bursts forth into your conscious awareness. The idea must finally be verified. Many of our ideas will turn out to be mistakes, but that's part of creativity. In the nearly hundred years of investigation since Wallas proposed this theorem, we have not moved much further forward in understanding creativity.

Alan Turing described his thoughts on the science behind creativity in a short piece he wrote about decision making:

> *"When making a decision of minor importance, I have always found it advantageous to consider all the pros and cons. In vital matters, however, such as the choice of a mate or a profession, the decision should come from the unconscious, from somewhere within ourselves. In the important decisions of personal life, we should be governed, I think, by the deep inner needs of our nature."*

Later in his career he came to believe machines would become intelligent and this sort of intuitive thinking could be effectively performed by a computer. As you know, I don't agree with his later viewpoint.

Another person who has thought long and hard about creativity is John Cleese, the comedian and actor. He describes the process wonderfully in a number of talks which you can find on YouTube. He finds a lot of his creativity emanates from his unconscious rather than conscious thought processes. To optimize this he needs large uninterrupted blocks of quiet



John Cleese on Creativity

time. Often, if a problem seems impossible, he will sleep on it. When he wakes the next morning, he will frequently find the problem has solved itself and a solution is ready at hand.

Art

The final class of creative thinking we generally recognize is artistic skill. This is probably a form of convergent thinking, except both the problem and the solution are open. Good artists are considered highly creative and most people tend to agree on what constitutes good art. There are some arguments but they are usually more about genres. I might not appreciate modern installation art, even to the point of declaring it, "not art." But, when forced to ignore their prejudices most people tend to agree on the distinction between good and bad.

Painting, sculpture, music, architecture and poetry are the traditional fine arts. There is often some argument over architecture: is it not too 'functional' to be considered an artistic endeavor? After all, art is not supposed to have any purpose other than to be, well, art. This definition inevitably leads to arguments about whether bad art is still 'art'. Art should be artful and how do you arrange a pile of used tires artfully? But this is a very narrow definition. I prefer to define art as something that provokes an emotional response in the beholder. Using this definition, the fact that a pile of tires disgusts and annoys you is exactly the point. Perhaps a more 'enlightened' viewer than you is intrigued by the clever use of materials.

Regardless, we consider art to be a creative endeavor and we can measure it using the criteria of novelty and quality. Since most people agree on these measures for a given piece of art, we can use the wisdom of crowds to give us a scientific scale. That does not mean there won't be art that you love but which leaves me cold. That is the joy of it. Novelty and quality are not the same as joy and pleasure, far less the tingle factor.

◆ ◆ ◆

A quick test for artistic skill is to take a pen and paper and turn to the person sitting on your left and try to sketch them.

TRY IT!

If you tried, you and your neighbor would probably find the results rather humorous. But, most likely, you did not follow my instruction. This is a form of social self-censorship. I asked you to do something rather difficult and embarrassing, but very creative and likely to enlighten you. Sadly most people – I am no exception – tend to censor their creativity for fear of embarrassment. Children, of course, do not sensor themselves as much as adults.

Now you know how to be more creative. Find your inner child and don't censor yourself too much!

What Sparks Creativity

As an inventor, I'm often asked what makes me creative. How do I do it? The answer is, I have no real process. After all, a process is mechanical and this entire book has been about exploring how creativity is a non-mechanical task. However, there are many things you can do to unleash your creative potential.

New ideas are often sparked through linking disparate ideas. Expose yourself to as many ideas as you can, read widely, attend conferences, visit customers.

Creativity requires peace and quiet. I personally get up early every morning. This gives me a good two hours of uninterrupted time every day. It's also the part of the day when my brain works best. Others prefer to work late into the night.

Pressure, for me and for many people, is a great incentive. Tales of the Polish Enigma code breakers, Douglas Adams' writing deadlines and the fear of impending death in shark attack stories all force people to think in an accelerated way. This appears to help many people defeat the human tendency to prevaricate.

On the other hand avoid panic. While a level of pressure can help, panic is unproductive. There is a sweet spot between having enough time to get properly acquainted with a problem and an impending deadline to force the crystallization of ideas. This balance varies from person to person and is something you need to test for yourself.

You need time off. Once you have a well thought out idea, you may need to leave it alone for a while to allow your subconscious to work. Time off does not need to be two weeks at the beach. Charles Darwin and Benjamin Britten used to go for long walks. You can walk in Darwin's footsteps at Down House in Kent. Others such as John Cleese like to 'sleep on it'. Stephen Hawking distracts himself by working on a different problem for a while. Anything that avoids focusing directly on the problem itself seems to allow our creative freewheel run.

Environment can be important. The campuses created by Steve Jobs for both Apple and Pixar are designed to foster creativity. The physical environments build team behavior but also cause people to bump into each other. Cross-pollination drives creativity.

There are also some myths to dispel about creative people. Inventors are portrayed as eccentric and hopelessly disorganized, but Feynman kept notes of every idea he ever had. I have kept a series of notebooks, now computer based, since I left university. I still have almost all of these on a shelf at home. Creative people may be a little mad, but the successful ones are rarely disorganized.

You must allow your brain to free wheel. J.K. Rowling has said the characters in the *Harry Potter* novels write themselves. I come to my computer each morning having *not* thought too much overnight and just write. Creation is just that; you must allow yourself to do it. It's not a process.

The Innovator's Dilemma

Why don't big companies create? In 1997, Clayton Christensen of Harvard Business School wrote *The Innovator's Dilemma*, the seminal work on creativity within organizations. In it, he shows us why established companies tend not to innovate and why startups exist. Christiansen's academic research examines how companies handle discontinuous change in technology, focussing on the hard disk industry.

You might not think this a very sexy sector. Microprocessors and game consoles would be more fun, but the great advantage with the hard disk is there is a single industry journal that has tracked the progress of every player over 30 years, collecting detailed annual data on every facet of their business. For an academic, this is gold dust.

IBM invented the hard disk drive in their research center near Winchester, England. The first prototypes were, consequently, called Winchester Drives. When Christiansen examined the industry, he found something very strange. As the size of disks reduced first from 8" to 5¼" and then from 5¼" to 3½", the dominant players in the previous era went out of business and new startups colonized the market. This fallout was not confined to a few small companies. It affected large, well-established, publicly-listed organizations, too. They failed en masse at each discontinuity. At first this made no sense to him. Surely a skilled

Hard Drives

hard drive manufacturer would be the obvious group to construct the next generation. But it seemed that not only did incumbent players *not* construct the next generation, they ran their businesses into the ground while ignoring the technology discontinuity. Despite their legions of Ivy League graduates and business school MBAs, they all went bankrupt.

As he looked around the economy, he found a similar pattern in other sectors. Minicomputer companies failed to make the jump to personal computers. Further back in time, buggy whip companies – in the Fortune 100 at the turn of the 20th century – failed to make the transition to the motor vehicle economy. The only exception he could find was IBM. IBM had successfully navigated some transitions but at that time was fighting for survival as companies transitioned from mainframes to Linux based servers and their survival was in question. Why was this so?

Christensen's conclusion is that established companies tend to concentrate too much on their existing revenue streams while ignoring potential new ones. This is no surprise. When the disc drive industry made the move from 8" disks to 5¼", the only customers for these new new smaller models were unheard of manufacturers of personal computers. Some were based in the dorms of MIT and Harvard – Dell and Compaq – not in the existing powerhouses of computing – Digital

Equipment and Wang. New technology often underperforms the existing forms. 5¼” drives were slower, less reliable, and cost more per bit than their 8” predecessors but, of course, in one respect they were better. They were smaller and lighter and could fit in portable computers. The new technologies did a different thing in a different way, and overcame their disadvantages later. This chain of events is repeated many times over: Yellow Pages overtaken by Google, Borders by Amazon, Blockbusters by Netflix. Disruptive innovation changes the rules of the game as well as the pieces in the game.

Christensen’s advice to companies is to separate your innovators from the existing business because their priorities will differ too greatly. Modern companies build entirely new divisions to create new products, or set up innovation labs to incubate ideas that would otherwise never get enough resources.

Reward for Creativity

There’s no doubt society values creativity very highly. One of the first tasks Thomas Jefferson undertook when he became President of the United States was to set up a patent system. He remained head of the patent office for over ten years. These days, the protection of creative

Harold Cohen, Computer Art

ideas through patents, copyright, and trade secret is big business and combines to form the practice of 'intellectual property'. Societies with the best protection of intellectual property are often the most successful. The USA is the unassailed leader, with Asian countries rapidly catching up. Poor old Europeans have struggled with an almost unworkable patent system for nearly 30 years; a genuine Europe-wide patent only came into effect in 2013.

Creativity in the economy is now extremely important, and nothing emphasizes the point more than the job market. During the 60s finding a job was easy. There was an almost unlimited range of mechanical jobs on offer. In the post-industrial age, almost all the mechanical jobs have gone. Today we need to be experts in a field, able to solve problems creatively. You can't expect to walk into a job and be profitably productive on the first day. Finding a job is harder and the cost of employing someone is greater.

Why did we Evolve Creativity?

Roger Penrose wonders why mathematical creativity evolved in humans since it only became useful in ancient Greece a few thousand years ago. He believes it must have been useful for something before this. But what?

David Deutsch thinks creativity developed to allow one human to understand the thoughts of a fellow human being. We can't precisely communicate the 'programs' we run in our heads. We are unable to download a detailed thought and put it on a memory stick. He thinks our creative capacity developed to help us pass skills from one to another. The ability to paint and sculpt is an accidental by-product of this adaption.

It's my view we evolved creativity to deal with new situations and puzzles in our daily lives. We use creative thought processes and ingenuity to come up with novel solutions for when we can't rely on programming or a store of rules. Otherwise, the very first unforeseen situation could kill us!

Computer Creativity

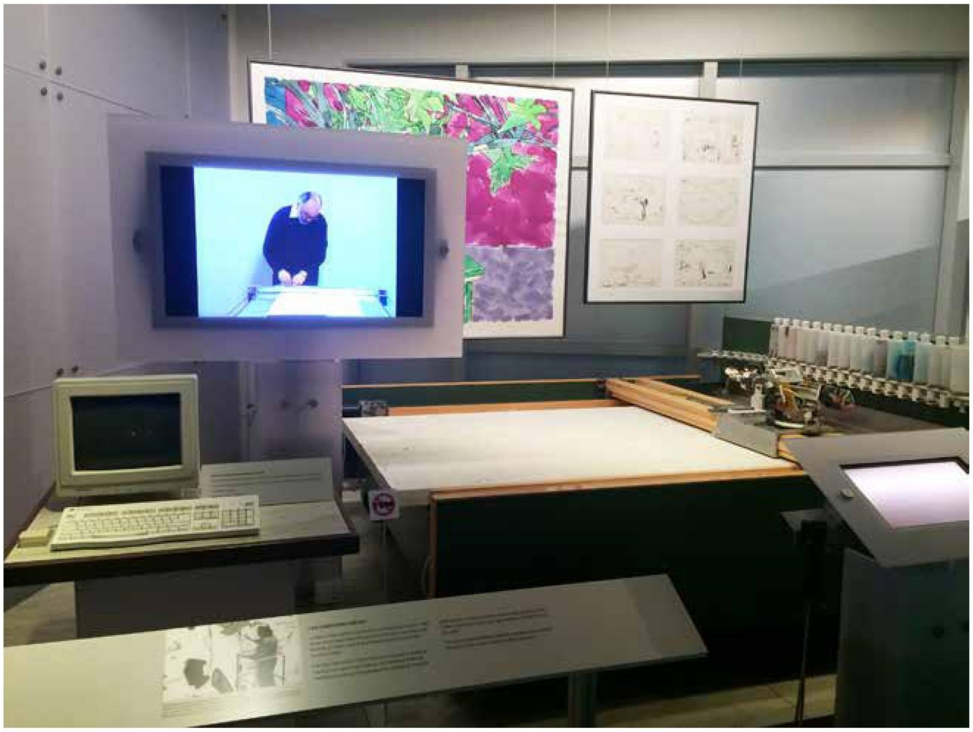
Humans find creativity difficult. It requires peace and quiet, detailed study and input of caffeine. How does a computer fare? I have argued that computers cannot be creative above the logic limit, so this does not preclude them from creating within the narrow confines of a particular solution space. But a human still needs to set the rules for this space. The level of creativity we should see from computers is convincing within a limited conceptual area. Computers are not going to wake up one morning and *decide* to compose a breathtakingly beautiful symphony, but if we give them rules they can make a convincing version.

Many computer systems have been designed to tackle creativity. We have already met the composer Emily Howell and Douglass Hofstadter's computation program. Here are two more examples: Jape and AARON, which create jokes and art, respectively.

Jape – Joke Analysis and Production Engine – is a program created by Graeme Ritchie and Kim Binsted. It generates puns, the sort of things you might find in an English Christmas cracker or children's joke book. I'll let the output speak for itself.

> Q: "What is the difference between leaves and a car?"
> A: "One you brush and rake, the other you rush and brake."
>
> Q: "What do you call a spicy missile?"
> A: "A hot shot!"

One of the most enlightening examples of computer creativity is AARON – refreshingly not an acronym. His machine is depicted here and you should look up some works on the web, such as *Adam and Eve*, and *Aaron's Garden*. The program encodes rules about figures, objects



AARON – Harold Cohen, Automatic Painting Machine

and perspective. Once coded the program takes off and is remarkably creative in its compositions, without any further human intervention. However, each new capability must be hand-coded by its creator, Harold Cohen. These paintings give a good visual interpretation for the sort of latitude imposed by the logic limit. AARON can do some very impressive things, but always in a mechanical – albeit beautiful – way, within the rules set by Harold Cohen, the true artist. AARON will not suddenly awake one morning and independently decide to experiment with the color blue!

To give an idea of what I mean by 'mechanical elaboration' in a musical context, imagine you were using a Casio synthesizer. These machines have all sorts of fun settings. You can program them to play drum tracks, fill in chords, and add a jazzy, syncopated harmonization to your melody. But none of this is truly innovative. It's mechanical elaboration of your artistic material. Jape, AARON and Emily Howell all do the same thing within their domain. They mechanically elaborate the artistic creation of their human masters. AARON does an extremely good job of this.

The Myth of the Design Tradeoff

An important consequence of the non-linearity of creativity is we are not constrained by tradeoff laws. Let me explain.

EMOTION
ECHNOLOGIES
Der ne
Polo

Volkswagen Polo

How often do you hear the phrase, "It's down to tradeoffs," or "It's a matter of priorities", or even "You can't have your cake and eat it."

These stock statements misunderstand the infinitely complex nature of creativity and problem solving. Let us take a concrete example: the car.

My first car was a Volkswagen Polo. It was a great little car, quite nippy, cassette-radio, and four seats. The most recent Polo has antilock brakes, airbags, NCAP 5-star crash resistance, smarter styling, and a low emission engine. Shall I go on…? The doctrine of tradeoffs says I would have to give up something to gain these new features. But, I have not. The newer Polo is cheaper in real terms than my original, as well as being better designed.

Problems always have at least two dimensions of freedom. We can trade one feature for another or we can innovate to both have our cake and eat it! We are never constrained to simple on-the-one-hand, on-the-other-hand type decisions. Creativity is unconstrained by linear rules and tradeoffs.

Process versus Creativity

How many times have you heard the words, "We must create a process for this!"

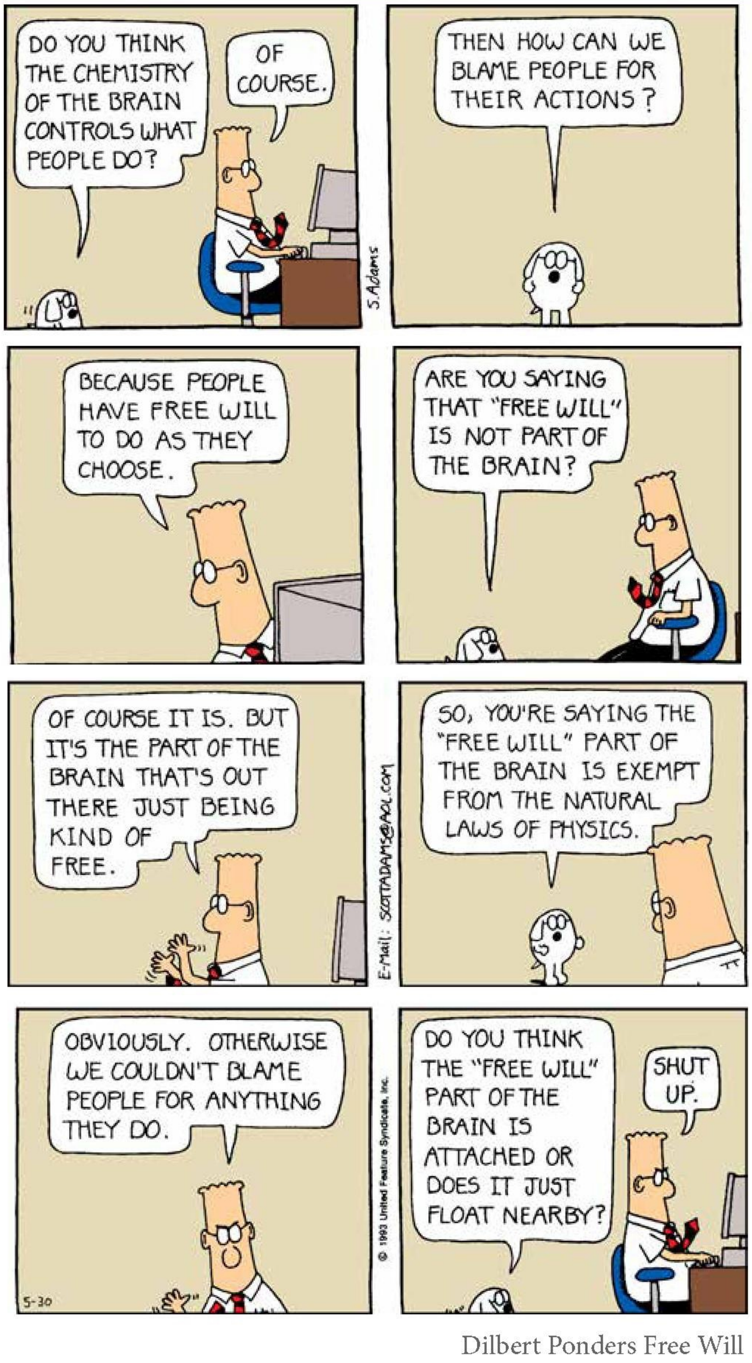
In its place, process is good; It makes things consistent, repeatable and predictable. You can follow a process by rote without error. Processes are also easy to document and communicate because they are symbolic. But process is limited. It is, after all, a set of prescribed rules for solving a particular problem – and, because of this, it falls into the same trap.

A process cannot solve a problem that requires creative thought or logic more complex than the logic limit. Logical processes are useful for tracking lists. I am reassured when I get on an airplane and know the pilot has been through a preflight checklist. I would not want to fly in an airplane where the pilot announced he was taking a creative approach to the preflight check.

Process is a perfect tool for organizing the steps around being creative, but it won't do the creating for you.

Chapter 15

FREE WILL


DO YOU THINK THE CHEMISTRY OF THE BRAIN CONTROLS WHAT PEOPLE DO?
OF COURSE.
S. Adams
THEN HOW CAN WE BLAME PEOPLE FOR THEIR ACTIONS?
BECAUSE PEOPLE HAVE FREE WILL TO DO AS THEY CHOOSE.
ARE YOU SAYING THAT "FREE WILL" IS NOT PART OF THE BRAIN?
OF COURSE IT IS. BUT IT'S THE PART OF THE BRAIN THAT'S OUT THERE JUST BEING KIND OF FREE.
E-Mail: SCOTTADAMS@AOL.COM
SO, YOU'RE SAYING THE "FREE WILL" PART OF THE BRAIN IS EXEMPT FROM THE NATURAL LAWS OF PHYSICS.
OBVIOUSLY. OTHERWISE WE COULDN'T BLAME PEOPLE FOR ANYTHING THEY DO.
5-30
© 1993 United Feature Syndicate, Inc.
DO YOU THINK THE "FREE WILL" PART OF THE BRAIN IS ATTACHED OR DOES IT JUST FLOAT NEARBY?
SHUT UP.

Dilbert Ponders Free Will

“We have to believe in free will -
we have no choice.”

Isaac Singer

“Time really is an illusion -
lunchtime doubly so.”

Douglas Adams

A child grows up in poverty, their father absent, mother a drug addict. Riots break out and the child defends the local convenience store. Another child born on the same road, but from a better background, loots the store and is arrested. This scene played out on the streets of London during the summer of 2011, but similar incidents happen all across the world. People choose different moral paths; one person makes a good decision; the other, a bad one. Did they make these decisions freely or was their behavior inevitable, dictated at the dawn of time?

Free will is at the heart of our justice system. It requires a crime to be intentionally committed by a person of sound mind. If I kill you in an accident or because I am mentally incapacitated, I am innocent. Of course, if I mentally incapacitate myself with alcohol I would be guilty of manslaughter, perhaps even murder.

Our justice system requires a crime to be intentionally committed by a person of sound mind. Whenever we see something bad in the world we trace the events back to the thought processes which led up to it. It seems we punish the decisions in our brains leading to a crime, not the crime itself. But, in a deterministic Universe my thoughts could never be at fault. They are inevitable. "The Universe made me do it!"

You need not worry about the fabric of society falling apart in a deterministic Universe. The whole of existence will play out according to a predetermined script, complete with lawyers, trials, drama and pathos. The judge, jury and executioner would also have no free will. It would look as if you paid the price for the choices you made, but this would be an illusion. The whole thing would be like one enormous screenplay.

The concept of determinism goes against our conscious experience. We all have a strong sense of free will. I certainly think I have it! And this presents a problem, because the classical laws of physics say our Universe is entirely deterministic, and that free will is an illusion.

I should briefly mention 'compatibilism', a branch of philosophy that claims determinism is *not* at odds with free will. It argues that if I feel free and my actions do not appear constrained, then I am free even though my future might be inevitable: a sensation of freedom is sufficient. This seems rather feeble. I am seeking an explanation for how we might be truly free to choose our actions, not some linguistic trick to argue freedom is subjective. I believe true free will is a physical principle with observable effects on the Universe that would not be seen in a determined one.

Domino Toppling

Determinism

To firmly grasp the idea of determinism let's look at a fun example, domino toppling. If I arrange a set of dominoes on their edge in a long line and push over the first it will fall, knock over the next, then the next, and so on until all the dominoes have fallen. It is inevitable, and fun to watch. The same is thought to happen with particles in our Universe, albeit at a much smaller scale. The laws of physics governing these particles describe precisely what will happen as they interact. Our Universe could be thought of as a mechanical clock, wound and set at the Big Bang, or a fractal equation generating the wonders we see around us.

When I push over the first domino it should be possible to capture all the information about the particles in the dominoes, my hand, the table, and the surrounding environment to precisely determine what will happen next. Will all the dominoes fall perfectly, or is there a break in the pattern – one domino just a tiny bit out of alignment – which will spoil the fun? All the information is there in front of me and I should be able to predict it perfectly.

The laws of physics, as we understand them, are not only deterministic, they are reversible. This means if we know the position and momentum of every particle in the dominoes and the surrounding environment, we can extrapolate their motion back into the past. It should be possible to trace back the path of each particle to reconstruct the past history of the dominoes.

If we were to cast a wide enough net, and collect all the available information, we could go back and see the events in the factory where the dominoes were made, or even see the trees that was felled to make them. We would need a *lot* of information and huge computing power, but we *could* do it! With a sufficiently powerful computer we could travel back in time, albeit as a simulation, and relive past events. This would have no effect on the events themselves as it would be like watching a movie, but we could see every aspect of the past from any viewpoint.

To perform this time travel trick for real, we would have to gather information from an enormously wide area. Information spreads out at the speed of light. One minute after the dominoes topple, the information about the event will have spread one light minute – that is over a sphere forty-million kilometers across – half way to Venus. An hour later and it would be outside the solar system. If you were to cast your net that wide, and gathered up all the information within the sphere, you could still perfectly model the moment when the dominoes were toppled. The wider you cast the net, the more information you have and the further backward in time you can travel. If you take the idea of collecting information to its logical conclusion you could gather all the information in the Universe at a moment in time.

In 1814, Laplace put this idea in an essay. He proposed an immensely powerful being observes the position and momentum of every particle in the Universe. Armed with *any* snapshot of the Universe and the laws of physics, the entire future and the past of the Universe. The being was nicknamed Laplace's Daemon and the idea has influenced philosophy ever since.

If the Universe is predictable, our concept of time needs to be rethought. A common sense notion is that things in the future are unknown and things in the past are known. But, in a deterministic Universe a daemon or a supercomputer could keep track of all the information and tell you what is inevitably going to happen. The conscious feeling we have of moving through time would be just an illusion. Past and future have no meaning and there would just be a solid, permanent block of space-time. If you stood outside the Universe and looked at this block of space-time, everything that is going to happen and has already happened is set. This is sometimes called the Block Universe Hypothesis and is the logical conclusion of any theory that imagines an entirely determined Universe.

One thing that seems to throw doubt on this Block Universe Hypothesis is our personal conscious experience of the world. We experience the Universe unfolding over time. (Of course, we could have this conscious experience in a determined Universe if someone had programmed it that way. All we can say is that it seems unlikely someone would go to the trouble of giving us a completely fictitious experience. There are an infinite number of possible Universes, why pick one where we think time flows, but it does not.)

Uncertainty

If you know a little of quantum mechanics you might imagine Heisenberg's Uncertainty Principle comes to our rescue.

Heisenberg's principle is often misunderstood. People sometimes try to explain it as an experimental problem. If I want to measure the position of a particle I am going to need to shine a light on it. The photons I use to illuminate the particle will knock it out of position so the act of measurement disturbs the system. This is *not* the Uncertainty Principle. It is a different but related effect, called the measurement problem. The muddle is really Heisenberg's own fault. When he tried to produce a layman's explanation he used the analogy of disturbing the particle with the photon. This is wrong. A photon would not disturb a particle enough to explain the uncertainty we find; particles are fundamentally uncertain even before we measure them. Heisenberg's Uncertainty Principle is a quantum property, which means it makes no sense and there is no analogy I can give you to properly explain it! Here is the closest thing I can find.

Imagine I am playing a musical note on a guitar. You might want to know two things about it; where exactly is the string and what pitch, or note, am I playing? The problem with these two measurements is they can't be stated at the same time. Pitch is dictated by the rate of oscillation over time: the number of times a string vibrates back and forth per second. Position is the exact location of the string at a given moment in time. If I state the position precisely this has no pitch because pitch needs a time interval. If I allow a time interval the string will move during that time and it won't be precisely in one place. The best I can say is the string is about two millimeters above the fret board and two-thirds of the way across it.

So, I hear you cry, this Uncertainty Principle means our Universe is not deterministic because it is uncertain.

Unfortunately, the principle only prevents *us* from measuring the position and momentum of a particle at the same time, it does not prevent the Universe *knowing* the information it needs to allow the particle to go about its business in an entirely deterministic fashion. There is a perfectly reliable and predictable wave function that governs the motion of every particle, just as there is an entirely predictable equation for the motion of a string on a musical instrument.

If both the classical and quantum laws of physics are deterministic where does the freedom come from to make our Universe non-deterministic? There is just one place to look: you and me.

The Observer

I am looking out of my office window. It is a sunny autumn day and I have a beautiful view over London, but if I squint a little I can also see my reflection. The window in front of me is not perfect. Although it is mostly transparent, the glass also reflects some of the light. If you think of light as particles, the majority of the photons go through. But some bounce back. I'm going to show you that the behavior of these photons is governed by the observer – me!

The laws governing light, and most of the strange and wonderful effects it has, were first stated by Isaac Newton. Newton was an extraordinary man. He discovered many of the physical principles we use today, and his view of the Universe reigned unchallenged until Einstein's discovery of relativity. He was also, by many accounts, a nasty piece of work. Not only was he a famous academic, he also head of the Royal Mint. He is said to have taken great pleasure in having forgers hanged on Tower Hill. He claimed the invention of differential calculus despite it being invented independently by Gottfried Leibniz. Newton managed to have himself appointed to chair the committee reviewing Leibniz's work and determine who had come up with the idea first. Unsurprisingly, the committee found for Newton!

We see Newton's laws of reflection and transmission in all manner of everyday products, for example, the antireflective coatings of camera lenses or the screen of your smartphone. Manufacturers cover the glass in these products with coatings just a few molecules thick. Interference between the layers *kills* the reflections. On a very expensive lens several different layers are used; some kill red light and others kill blue light. Together they suppress most of the reflection. If it were not for these coatings you would be unable to go to the park on a summer's day and read your iPad. We need to think about reflection and transmission to demonstrate our role as observers.

A windowpane has two surfaces. Both surfaces reflect light, and if you look closely you will see your face is really reflected twice. You might think this is simply a double reflection but this is not so. Light behaves like waves. As with water waves, they interfere with one another. If two light waves are at the top of a crest as they meet, the result is a crest of double height. If both are at the bottom, you have a double trough, and if one is a crest and the other a trough you get nothing as they cancel each other out. You can see this effect in waves on the surface of a pond.

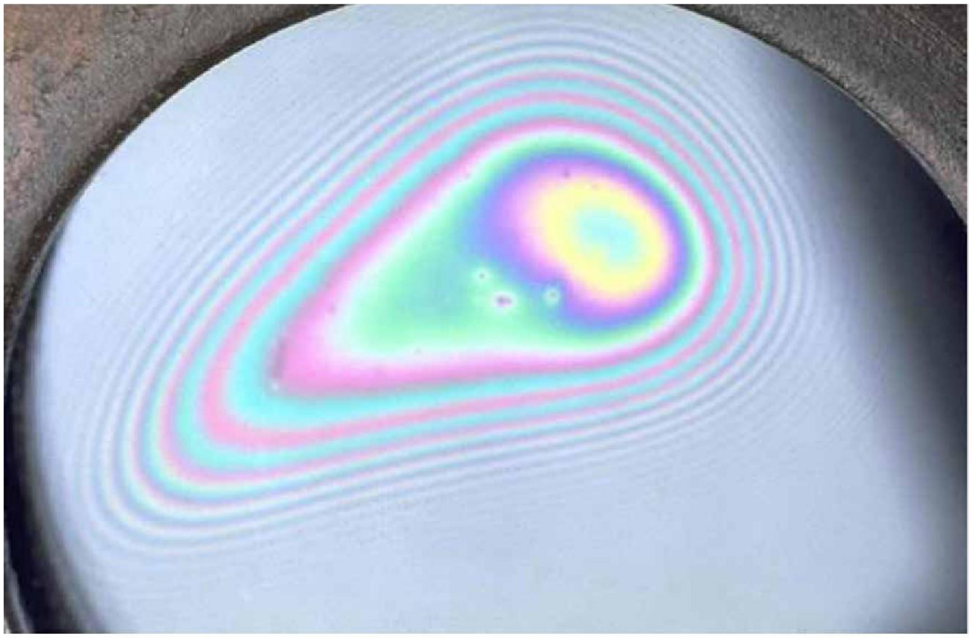
When light strikes a window pane, the light has two chances to reflect: one from the front surface and the second from the back surface. These two reflections interfere with each other. And, again, when you

Wave Interference

have interference you sometimes get constructive interference – the double crest or trough – and sometimes destructive interference – the crest and trough canceling each other.

The reason we don't see this effect in every reflection is because of imperfection. Windows are not perfectly flat and light is multicolored, so the effect is hard to see. But it is definitely there, and if you look really hard at a reflection in a window, you can sometimes see it at the boundary of sharp objects. The effect is very clear in the next picture which has been set up with two flat pieces of glass resting against each other. There is a tiny air gap between them. It is also commonly seen on puddles in cities. In this case the puddle usually has a slight film of oil on it. Unlike glass, puddles are perfectly flat thanks to gravity, but the oil film is thick in the center and thin at the edges. The light reflecting off a puddle will show a rainbow of colors. This is because each color is giving us a pattern of light and dark. If you look at the same puddle at night in a yellow streetlight, you will just see a monochrome pattern of light and dark. Look at the picture of waves on a pond.

The patterns of light and dark are usually explained by imagining the photons interfere with each other. This explanation is insufficient. Imagine for a moment you can see as well as a frog, and perceive one photon at a time. One moonless night just one photon comes to the two glass surfaces and reflects. What happens?



Newton's Rings

It turns out a single photon can interfere with itself! How can this be? It must somehow split up and consider both the available paths reflecting off the glass surfaces.

Now that we have these two concepts in our head, that a photon sometimes reflects and sometimes does not, and a single photon must consider both paths, we can ask: what tells the photon what to do?

There are only three possible answers: the light source that emitted the photon, the pane of glass that reflected the photon or the observer that saw the photon – me.

The first obvious place that might control the photon is the original source of the photon; the light bulb. The photon might leave the bulb already knowing what path to follow, whether it will be reflected and whether that reflection will be affected by the gap between the two surfaces in a positive or negative way. This is sometimes called a pilot wave theory. The problem with this theory is I could insert a piece of glass into the experiment after the photon has left the light source. This will affect the photon but the light source could not have known my intention in advance and told the photon what to do. Therefore, the path of the photon is not pre-programmed by the light source.

Now our photon has left the bulb and is traveling toward the glass. The glass has two surfaces. The photon reaches the first surface and has to decide if it will reflect. But there is a problem. The second surface will have an effect on this decision – constructively or destructively. The photon can't make up its mind at the first surface. It has not yet seen the second surface.

The photon travels on and reaches the second surface. It needs to make a decision: Shall I reflect or not? It cannot decide that it should have been reflected from the first surface because it is already at the second surface; it's too late.

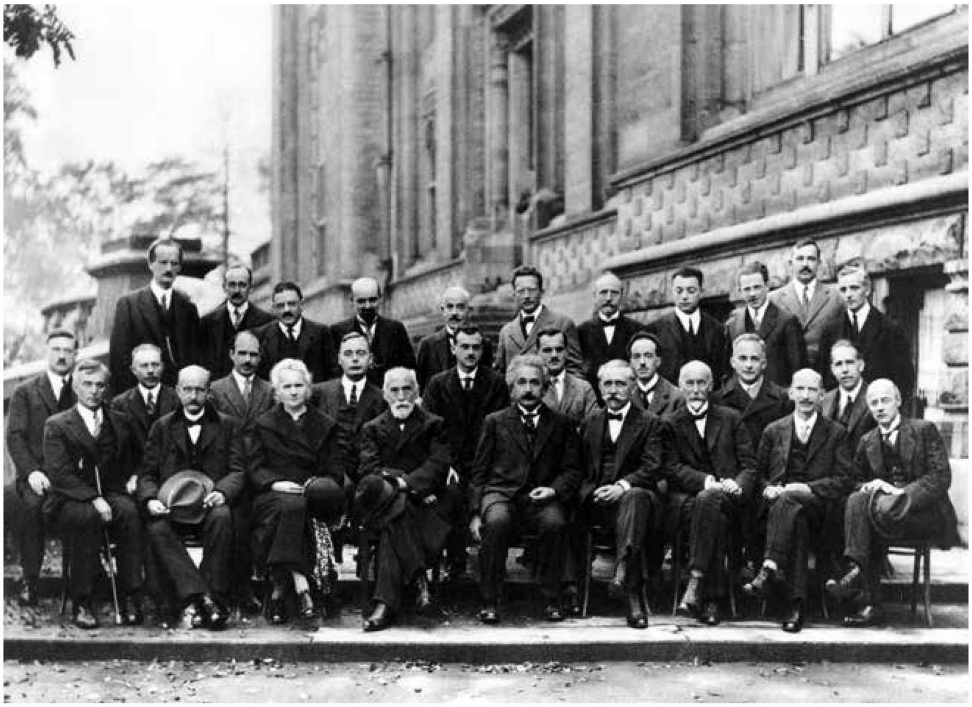
The photon is stuck. It cannot make the decision at the first surface because that is too early, nor at the second surface because that is too late. The glass surfaces cannot be the source of the decision.

This leaves only one remaining option: I, the observer, *tell* the photon what to do. The word 'tell' is probably a little strong. Sadly, I am not that powerful. All I can do is tell the photon to make up its mind. When the photon reaches my eye, it *must* decide what happened to it along the journey, but this decision appears purely random and I have no effect upon it. The best way physicists have found to describe what is going on is to say particles, such as a photons, behave according to a wave function. Particles oscillate between all the possible options available to them and when we take a measurement this freezes the oscillations and gives a single result.

Where exactly is the measurement made? At my eye when the photon is refracted by the lens, when the photon enters the aqueous humor, or perhaps as it interacts with the rods and cones in the retina. Maybe we must wait until the detection of the photon is converted into an electrical impulse in the optic nerve or even the point at which my human consciousness perceives it.

This prompted the physicist John Bell to ask a slightly tongue-in-cheek question, "Was the world wave function waiting to jump for thousands of millions of years until single-celled creatures appeared? Or, did it have to wait a little longer for some more highly qualified measurer – with a Ph.D.?" You see his point. Where is the bar set that defines a measurement?

One of the most extreme answers to Bell's question is the strong anthropic principle. It argues humans – or at least sentient beings, perhaps even cats – *cause* the Universe to exist. The Universe bubbles along in a state of superposition with every possible event occurring and bifurcating until an observer emerges in one of the branches and the whole edifice collapses to that state. It is not clear if this produces many concurrent universes or if the first universe with a sentient being wins!



Solvay Conference

"There is no way to understand the mechanism that turns the water of the brain to the wine of consciousness."

Colin McGinn

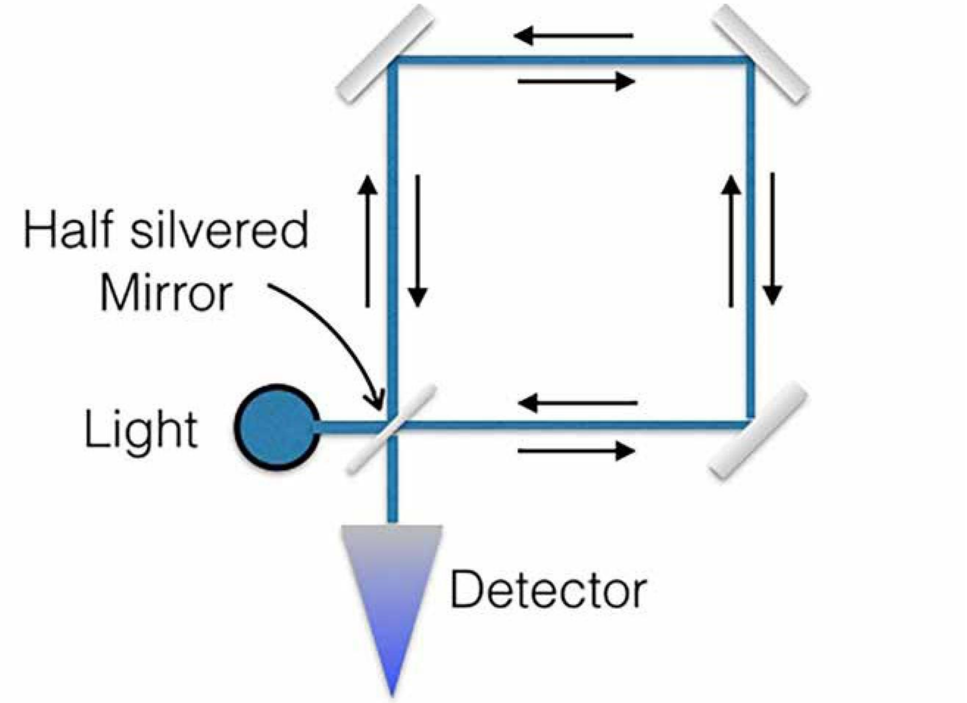
"If you think this Universe is bad, you should see some of the others."

Philip K. Dick

Schrödinger's Cat

Measurement is a big puzzle. What causes the collapse of the wave function so that the photon stops considering many optional paths and makes a hard and fast decision. When light passes through a series of glass surfaces, it is reflected or transmitted by each. We can stack up as many pieces of glass as we want, but none of the surfaces will cause a measurement – a collapse of the wave function. It is not until the photon reaches a detector that a measurement is made and all the potential reflections and transmissions that might have happened 'collapse' into the one choice that actually happened.

You might doubt this but there are ingenious experiments that can be performed to prove it. One is quite simple to do and can be set up on your kitchen table with a handheld laser and $100 worth of optical components. You need three ordinary mirrors and a beam splitter. Beam splitters are often made from half-silvered mirrors. They are similar to your bedroom mirror except the silver coating is more thinly applied, allowing only half the light to reflect while the rest passes straight through. Arrange the mirrors and beam splitter on a table at the four corners of an imaginary box, as in the diagram. If you point your laser at the half-silvered mirror, half the light will go straight through and half will be reflected upwards towards the first mirror. It is sent on around the square until it meets the half-silvered mirror again, and the beams meet up. You might expect that half the light reaches the detector but this is not what happens. Depending on the way the mirrors are positioned, either all the light reaches the detector, or none does. (The light does not disappear it just gets sent back to the light source, energy is conserved.) If you turn down the brightness of your laser, this does not change. Even

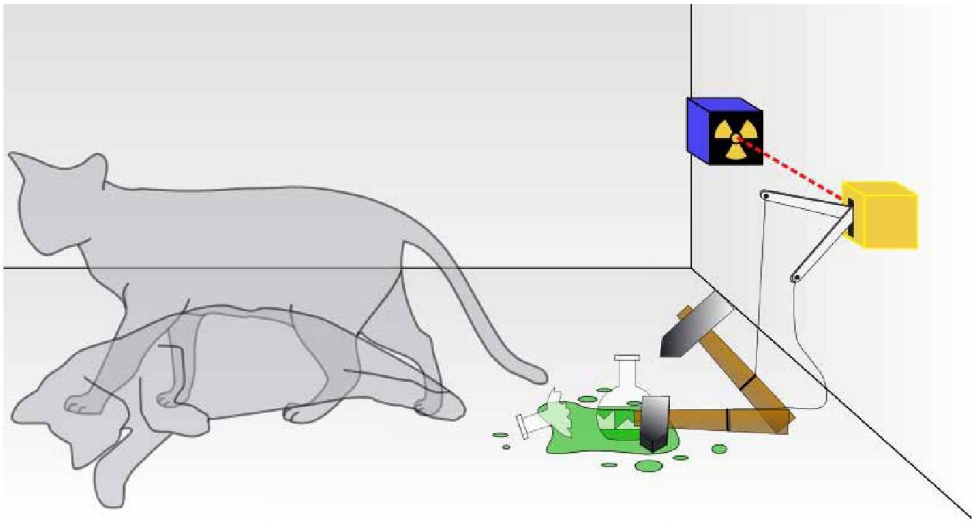


Interferometer

if only one photon is traveling at a time, still no light comes out in one direction and all the photons come out the other. The wave functions of each photon interfere with each other constructively or destructively. The only conclusion available is the photon must be traveling along both paths! They are said to be in 'superposition'. If you introduce a measuring device half way around the experiment, it will destroy the superposition and the photons behave in the common sense way. Remove the measuring device and, once again, the photons seem to take both paths. Richard Feynman pointed out that you really have to imagine that the photons take every possible path, not only the straight line paths. He received his Nobel Prize for demonstrating how to add up these infinite paths to get a finite answer with his 'sum over histories method'. Superposition is a strange idea when limited to the realm of small particles but what about larger things? – cats for example.

Erwin Schrödinger's unfortunate cat is trotted out to demonstrate the paradox so often that Stephen Hawking is on record for wanting to reach for a gun every time he hears mention of it.

The thought experiment works like this. A cat is put in a box with a radioactive substance, a Geiger counter and a vial of poison. If the counter detects a radioactive decay it breaks the bottle and the cat dies, if no decay is detected the cat lives.



Schrödinger's Cat – both Alive and Dead

Since radioactive decay is a quantum event, we have to assume it might or might not have happened right up until the point of measurement. It is the same with any quantum event: photons reflecting from a piece of glass, measuring the spin of an electron or measuring the polarization of a photon. All these quantum effects exist in superposition until measured. But in the real world, we don't experience superposition. If I miss the train, I miss it. I don't partially catch it and partially miss it, and I don't experience any such quantum ambiguity. The only place I ever see such effects is watching science fiction movies. In real life the large scale world is certain. At what point does this quantum uncertainty transition to our classical certainty? What is the state of the cat before I – a sentient observer – open the box? Was the result of the decay measured by the Geiger counter, the cat, or are we waiting for someone to open the box and observe the result?

The Copenhagen interpretation of quantum mechanics – named after the main center of early quantum theory at the Niels Bohr Institute – says the cat is both alive and dead until I make a measurement. The cat is said to be in superposition, meaning a live cat and a dead cat inhabit the same volume of space-time 'experiencing' both alternatives and waiting for my measurement. This seems nonsensical, but Copenhagen quantum folk simply say, "That's the way it is; the mathematics works, if you don't like it, tough. Nature does not have to explain herself."

Einstein strongly disagreed with this position. He believed the world is certain and laws must govern radioactive decay and, therefore, the breaking of the vial and the life and death of the cat. There must be some, as yet, undiscovered theory. He reasoned as follows: A particle has position, velocity and spin. Why can it not have more hidden information that tells it when to decay? Perhaps particles are composed of sub-particles that cause the weird quantum effects we see.

We *have* discovered sub-particles – quarks and the like – but more than a hundred years of experimentation have gradually ruled out any form of theory explaining how these random events can be governed by the properties of the particle. The collapse of the wave function just seems to happen randomly.

There is one explanation for quantum mechanics that avoids the measurement problem altogether but it is even stranger than the Copenhagen interpretation: 'the many worlds' view'. The idea was first put forward by Hugh Everett in 1957, and it claims measurements are never made, there is never a collapse of the wave function, and every wave continues to exist. We just can't see them all. There is a version of me that has seen a live cat and another in a parallel universe that saw a dead one. The two versions of me are also in superposition, just like the cat, so there are an infinity of parallel universes tracking every possible option.

The only measurable consequence of this 'many worlds' idea is the existence of enormously enjoyable science fiction plots and much poking of fun between the many worlds camp, and the no-many-worlds camp. The single-worlds proponents point out the whole idea is untestable and just plain odd. For example, each choice we make, every reflection and any quantum process generates a new branch in the Universe. This is a vast amount of information to track and puts us back in a position where moral choices have no consequence. Every decision I make spawns a Universe where I made a different choice.

For a humorous take on this, you can visit a website and buy your own Universe for $2.99. You pose a question based on the throw of a die, let's say, one to three I go to work today and, four to six, I take a sick day. The website generates a quantum random number using an experimental setup at a laboratory in California. You can make your choice based on this quantum random number in the certain knowledge that another Universe springs into existence where you made the alternate choice, so somewhere you are not taking a sick day after all, and can be found hard at work at your desk.

There is one more explanation for quantum measurement, proposed by Roger Penrose. He proposes gravity comes to the rescue. Once enough particles are involved in the superposition of states, the curvature in space time becomes great enough to force a measurement event. In his view, a measuring instrument is simply an amplifier which brings a quantum event to the point where gravity begins to matter. His solution removes the requirement for many worlds and, indeed, our curious position as the conscious beings that bring the world into existence. Of course, Penrose does not stop there. He proposes the quantum gravity interaction gives rise to conscious thought and this process is the root of mathematical intuition.

The aim of our discussion is to show where determinism might break down in the physical laws of the Universe. If our Universe is determined, there will need to be a huge quantity of information stored somewhere to tell it what to do at each step. Storing a script for the Universe is not the conventional way people imagine determinism works. Rather they explain the apparent complexity we see through the application of a simple set of rules called 'the laws of physics'. We imagine using these laws to expand up a small set of starting conditions into the complexity of the Universe we experience. This is similar to the way fractals produce beautifully complex images from a tiny quantity of information. For example, the Mandelbrot set is created from a single, simple mathematical statement just twelve characters long, with one or two starting numbers. If this is how our Universe works then our thoughts and actions are just like the fronds of a fractal. It would mean the particles in the Universe 'know' what they will do next and carry enough information with them to determine their next action. This is a testable hypothesis and the test we have devised to measure this is the twin particle experiment.

Right and Left Socks

> *"Was the world wave-function waiting to jump for thousands of millions of years until single-celled creatures appeared? Or did it have to wait a little longer for some more highly qualified measurer - with a Ph.D.?"*
>
> John Bell

Twins

There are several ways to make twin particles. The 'easy' way is with a laser and a nonlinear crystal. A beam of ultraviolet photons enters the crystal and about one in a billion times they interact with quantum fluctuations in the crystal lattice to create two red photons. This is known as 'spontaneous down conversion'.

There are two types of down conversion. In a type I interaction the twins have the same polarization, and in a type II they are at 90 degrees to each other. You can set up the experiment with either type, but it is important to remember which you used, or you will easily become confused. When we *talk* about photon experiments, we are usually referring to type I interactions because they are easier to understand. Often the actual experiment uses oppositely polarized photons because they are easier to generate.

Polarization is a wavelike property of photons. You can visualize them wiggling up and down, side-to-side or something in between. We use polarization to our advantage when we go on holiday to the beach; light from the sun is randomly polarized, but when it glances off the ocean it becomes predominantly horizontally polarized. If we wear vertically polarized sunglasses the glare off the ocean is blocked and we can see the ocean more clearly. The following two pictures show this effect.

The two photons we make with the crystal can be separated and sent to different places. The record so far is two towns near Geneva, 50 kilometers apart. For this experiment scientists 'borrowed' the unused fibers of Swisscom in the middle of the night – when phone traffic was light. A detector was placed at the end of each fiber to measure the particles.

Effect of Polaroid Lenses

When scientists examine the polarization of these photons, they get random results. Sometimes the photon is oscillating side to side, sometimes up and down and sometimes part way in between. This can be determined simply by taking a lens out of a pair of Polaroid glasses, holding it up at an angle and seeing if the photon can pass through. Obviously laboratory grade Polaroid material is available, so scientists don't have to destroy an expensive pair of designer glasses, but the principle is identical.

Very strange things happen when the measurements are made. The polarizations appear to have no discernible pattern, but once one of the photons has gone through a polarizer in the first town, its sister photon will always be found to have the opposite polarization (or the same if it was a type I)..

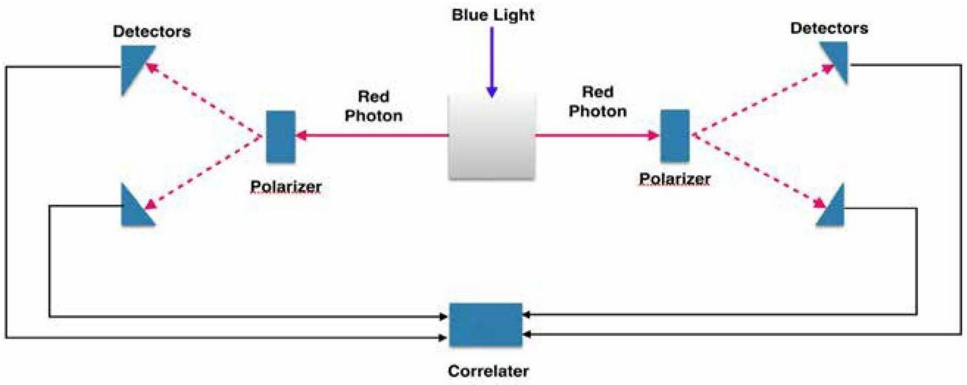
Einstein was uncomfortable with this for two reasons. The first related to his famous statement, "God does not play dice with the Universe." He was deeply uncomfortable with the idea that the polarizations were random. Even more troubling to him was the idea that the sister photon somehow instantaneously had the opposite polarization. How would it know? For the sister photon to immediately have the opposite polarization, information would have to travel faster than the speed of light from the first photon to tell its sister what to do. In 1935, Einstein wrote a paper with Jacob Podolsky and Samuel Rosen describing this 'EPR' paradox. Since faster than light communication was impossible – it breaks the law of special relativity – they concluded quantum mechanics must be wrong, or at least incomplete. A deeper theory would be needed to explain the particles' behavior. One very simple explanation is by analogy to socks! (Clothing analogies are one of the ways physicists try to make quantum mechanics less intimidating.)

Consider sister photons as if they were right and left socks. If we found a left sock on the bedroom floor, we would be unsurprised to find the matching sock was a right one. There is no need for messages to flow between the socks faster than the speed of light to synchronize them, they already know what they are! Einstein presumed sister photons were like socks; they were emitted from the light source with their polarizations already set, though you could not see this information until you measured one of the photons. The information was dubbed 'hidden' and the theory is called hidden variable theory.

Einstein was to be proven wrong.

For many years after the EPR paper was published, physicists split into factions: some thought the world random, some believed in hidden variables, and others thought attempts to 'understand' quantum mechanics were misguided. Why should physics make sense? The equations work. Who cares why?

In 1964, John Bell, an Irish physicist working at the Conseil Européen pour la Recherche Nucléaire ('CERN'), devised a way to test Einstein's hidden variable theory. He pointed out that if photons possess hidden variables and we randomly measure them with a detector set at three angles, we would expect to see more than one-third of the photons share the same result. But, in 1972, Freedman and Clauser performed this experiment and showed the photons share the result only a little over a quarter of the time. Since 'a little over a quarter' is less than 'more than a third', Bell's theory is false. Of course, Bell was entirely happy about this, since he set the equation up to be disproven. His equation is called an inequality because the equation contains a more than sign '>' rather than an equals '=' sign, so people say that quantum mechanics violates the Bell inequality. Because the inequality is violated, photons can have no prior knowledge of their polarization.



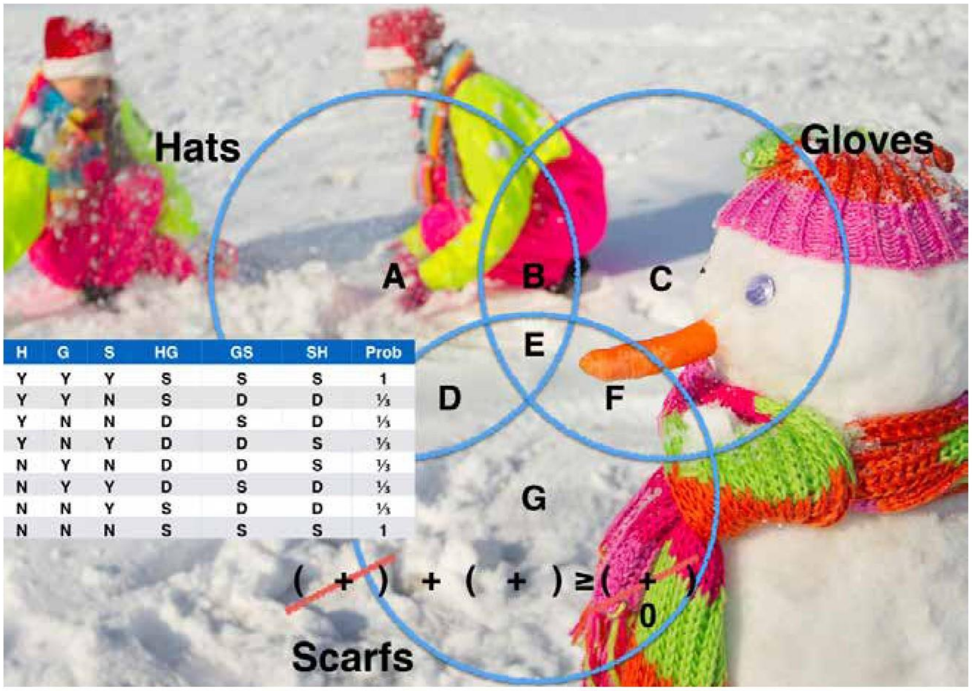
Bell Test Experiment

This is quite a complex piece of mathematics so let me show you how it works. Again, our thought experiment relies on an analogy involving clothing – sorry.

In the Bell Test experiment three polarizers are set up at 0, ⅓ and ⅔ of the way around a circle, 120 degrees apart. For Einstein to be correct photons must each carry at least three pieces of information:

If I meet the 0 degree polarizer do I go through or not?
If I meet the 120 degree polarizer do I go through or not?
If I meet the 240 degree polarizer do I go through or not?

If a photon had only one piece of information, say that it was vertically polarized, it would not *know* what to do if it came across a polarizer at 45 degrees. In that case the photon would sometimes go through and sometimes not, with a fifty-fifty probability. But Einstein did not want to countenance probability. "God does not play dice with the Universe." He required certainty. "I like to think the Moon is there when I am not watching it." The photons must know enough to handle, with certainty, any eventuality they may come across. (We could set up experiments with a more complex set of choices, dividing the photons into quarters, fifths and so on, but thirds are simple numbers and we can use the children's clothing analogy to demonstrate the mathematics.)



H	G	S	HG	GS	SH	Prob
Y	Y	Y	S	S	S	1
Y	Y	N	S	D	D	⅓
Y	N	N	D	S	D	⅓
Y	N	Y	D	D	S	⅓
N	Y	N	D	D	S	⅓
N	Y	Y	D	S	D	⅓
N	N	Y	S	D	D	⅓
N	N	N	S	S	S	1

Hats, Scarves, and Gloves

We could liken photons knowing three pieces of information to children in a playground choosing to wear either hats, scarfs or gloves in some combination: hats for vertical, gloves for 120 degrees and scarfs for 240 degrees. There are eight choices for each child; nothing, hat, gloves, scarf, hat and scarf, hat and gloves, scarf and gloves, or all three.

Bell asked how often we would see two measurements agree. Look at the illustration and you can see when this happens. If a child was wearing all the clothes then if you check any pair, say gloves and scarfs you will always get a yes. If one of the children is wearing none of their winter clothes, you will always get a no for any pair you check. In these two cases, we are always sure to get agreement. For all the other cases, only one in three of the tests will agree. So Bell said that any time you have something with three hidden variables, there is at least a one in three chance that the measurements you make will agree, since six of the tests are one in three and the other two are certain.

Due to Heisenberg's Uncertainty Principle we can only look at one piece of clothing at a time. But, there is a trick. If there are identical twins among the children – who always dress the same way in our analogy – we can look at the gloves of one twin and the hat of another. Because they are twins if the first twin is wearing a hat we know the second one is too, without looking. We have a trick to measure two things at once.

When the test is done on twin photons only one in four, one quarter, agree. So there is a problem with the children analogy. It turns out photons don't wear gloves, hats, and scarfs. There are no hidden variables. A photon does not know what it will do before you measure it and can only decide on the fly at the point of measurement.

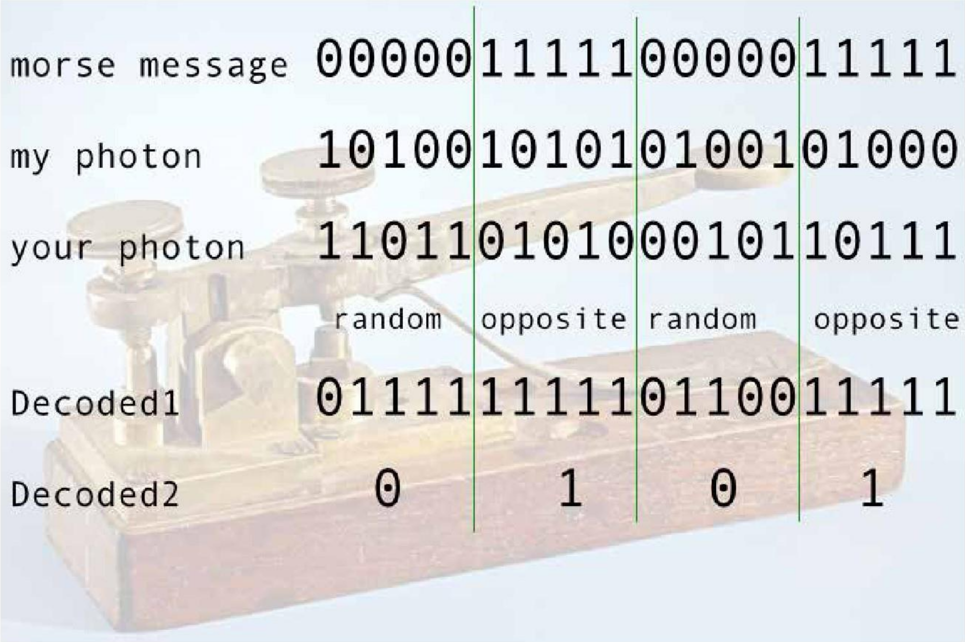
This means quantum particles are not *there* when they are not being observed. Observing them *does* appear to make them real. If the hidden variables, the gloves, hats and scarfs were in set positions when we were not observing them, the photon measurements would agree at least one-third of the time, but they do not. When we measure them, the two particles somehow communicate and agree to give a positive result only one quarter of the time. Bizarre, but that's just the way it is!

The Bell result is still somewhat controversial and has not been proven to everyone's satisfaction. Potential loopholes exist but are steadily being eroded. An experiment by Nicolas Giseng of CERN using the fiber optic network of Swiss Telecom to separate twin photons, shows the coordination signals must travel at least 10,000 times the speed of light – the limitation, and reason it is not infinitely fast, being the accuracy of his clocks. Daniel Sego, Daniel Danziger and Michael Wise have performed the Bell test experiment with an apparatus installed near Innsbruck where the choice of detector orientation was made by a random number generator after the photons had left the emitter. This shows the photons really can't know what they will do before they start their journey. Another loophole is the loss of some photons. We don't measure *all* the photons in an optical experiment because some are absorbed by the apparatus. It has been suggested all the 'lost' photons make up the error in the experiment. This is not very likely, it's akin to assuming all the voters who did not vote in an election would have voted Democrat. To avoid this criticism, an experiment has been performed with magnetized particles that don't get lost. The Bell result holds true.The loopholes are diminishing and it seems likely Bell will win out in the end.

Although the coordination information appears instantaneous, John Bell gave us an elegant explanation as to why this does not allow us to use the effect to transfer information faster than the speed of light. Imagine we are sitting at opposite ends of a room. We both toss coins and each of us write down our results; heads, tails, heads, heads and so on. I then acquire a magical power that causes your coin to make an extra flip just before you catch it, so it always gives the opposite result to mine. Although I am now controlling your coin, you cannot tell, as the result looks as random as before. The difference is simply that at first the coin orientation was random in its own right and then the opposite of my random result. It is only when I walk over and compare our results we can see they are matched in this strange way. There is no way to transmit information using this effect. Only after the experiment is finished can we exchange the necessary information to see the coordination that existed, and that comparison required me to transfer information. The fastest way to do that is at the speed of light.

Despite saying it is impossible, let us do a thought experiment and try to transmit information using Morse code. I will set up a simple old fashion telegraph machine. When I press the telegraph button at my end this will cause a measurement and the photons at your end will be forced to the opposite polarization. When I lift the key, your photons will revert to being randomized. You can see this illustrated in the diagram. Although I make your photon take up a polarization, analogous to making the coin take an extra flip, you don't have enough information to know this.

Now we are ready to use our quantum Morse machine to prove the Universe must have free will, or at least a degree of non-determinism.



morse message	00000	11111	00000	11111
my photon	10100	10101	01001	01000
your photon	11011	01010	00101	10111
	random	opposite	random	opposite
Decoded1	01111	11111	01100	11111
Decoded2	0	1	0	1

Quantum Morse Machine

A Simple Free Will Theorem

In the quantum Morse machine, I *do* transmit information faster than the speed of light. But the information I have transmitted is useless as it is, in effect, encrypted using a one-time pad. The only person in possession of a copy of this one-time pad is me: the sender.

Claude Shannon proved a one-time pad is unbreakable during the Second World War. Yet the British succeeded in breaking it. How was this possible?

The fatal weakness in the German one-time pads was the random numbers used to code the messages were generated by a machine, and were therefore not truly random. The numbers followed a sequence, and it was possible for Allied code breakers to work out the sequence and decode the messages. It follows that if we believe no message can propagate faster than the speed of light, my sequence of numbers must be non-computable. There must be no algorithm or computation that could generate it. Otherwise it would be liable to the same sort of decryption attack that the one-time pads suffered. If sequences of random measurements taken in the universe are non-computable it follows the Universe as a whole must be non-computable.

There are a few holes you could pick in this argument. Would it be sufficient if it were impossible to decrypt the message in the age of the Universe? What if there was an algorithm, but it was practically unknowable? But, I am talking here of principle. In principle, the Universe must be non-decryptable.

THERE'S PROBABLY NO GOD.
NOW STOP WORRYING AND ENJOY YOUR LIFE.

Richard Dawkins and the Atheist Tour Bus

"God exists, if only in the form of a meme with high survival value, or infective power, in the environment provided by human culture."

Richard Dawkins

Does God have Free Will?

Any discussion of free will is incomplete without some mention of God. Scientists generally avoid the topic, but since we're talking about such a fundamental concept, we must consider whether the Universe would be any different if it had a creator.

Recently there have been two widely publicized attacks on religious belief from the scientific community: the head-on attack from Richard Dawkins in *The God Delusion* or, the hard hitting sideswipe by Stephen Hawking in *The Universe in a Nutshell*.

Hawking made the front pages in 2000 with the statement: "There is then no need for a creator." He was considering whether God needed to ignite the Big Bang or if it occurred as a natural result of the laws of physics. Hawking had run the mathematics and realized a god was not needed to light the blue touch paper for the Big Bang - the laws of physics spontaneously caused it. His argument does not actually preclude the existence of a god, but it does move the point where we need a creator one step further up the chain.

This is not a fundamental change to the progress of theological argument over the last thousand years. Once we abandon our vision of God as a master builder, literally breathing life into Adam while putting the finishing touches to the Garden of Eden, we can move him up the causal chain as far as we like, eventually reaching a point where intervention is necessary to get things started. Hawking is only pointing out an intervention is not needed at the point of the Big Bang. It still begs the question “Where did the laws of physics come from?” If you don’t believe in a god then pushing a creator figure further and further up the chain eventually makes him redundant. If you have faith, you can take the position God is the creator of the fundamental rules.

Regardless of your personal position, I would like to make the argument for free will independent of belief. We must resolve the age-old paradox: How can God be all-knowing and all-powerful, and still have free will?

This is a long-standing theological debate dating back to the 15th century. It splits theologians into two camps. The first maintains God has both omniscience and omnipotence, and they are not inconsistent. This is the compatibilism argument again. Despite the acknowledged paradox, they argue that we should simply accept it and acknowledge that we are unable to comprehend such things. I don’t like this argument because it essentially denies reason. We are supposed to acknowledge that we simply cannot understand the mind of God. I prefer the more modern argument from the second camp that omnipotence trumps omniscience. It preserves the view that man can reason about the Universe – “Man is made in God’s image.”

This argument follows the logic: God must be able to choose not to know what will happen in the future so that he can have free will.

Fork in the Road

"When making a decision of minor importance, I have always found it advantageous to consider all the pros and cons. In vital matters, however, such as the choice of a mate or a profession, the decision should come from the unconscious, from somewhere within ourselves. In the important decisions of personal life, we should be governed, I think, by the deep inner needs of our nature."

Alan Turing

The Free Will Theorem

In 2006, John Conway and Simon Kochen published *The Free Will Theorem*. The paper received huge press attention and has been widely discussed in the scientific community. Their theorem states that; provided experimenters are free to run their experiment as they choose, the behavior of the particles they experiment upon is not determined in advance. Particles have free will!

If we go back to the Bell Test experiment, this proved twin particles do not carry around a parcel of information telling them what to do. Perhaps they get their marching orders from some outside influence. There are two possibilities. A particle is either told what to do by its environment or it gets its information from some data source. Can we use the laws of physics to test these possibilities? We don't need to know how the influence works, just that it might exist in principle.

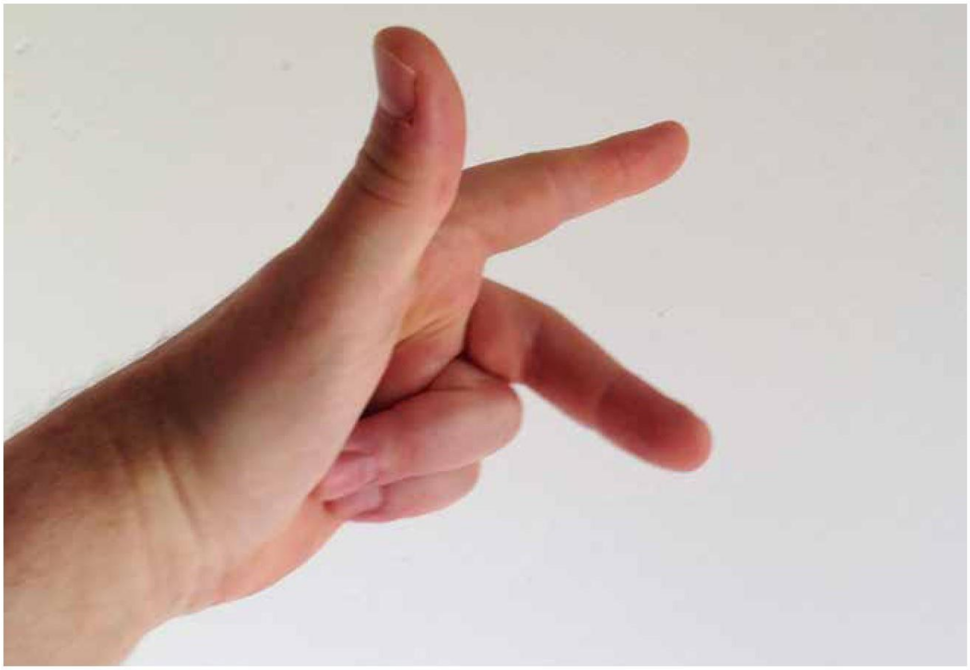
Conway and Kochen prove there can be no external influence, and when a particle reveals its spin, that spin was not known beforehand. It is independent of any information in the history of the Universe up to that point.

Conway and Kochen's proof is elegant and involves some mental gymnastics, but it is no harder than Archimedes' proof of the infinity of primes we looked at earlier. Let us start with our twin particles. We are going to pick bosons, which have whole number spin. If you measure the spin, you will always get a reading of -1, 0 or +1. 'Spin' is one of those words physicists use to explain quantum things. It does not necessarily denote rotation but, if your mental model is a spinning top, that's not too bad. If we measure something in the quantum world, it always yields a classical result – in this case the magnitudes of spin are 1, 0 and 1. (Minus one squared is one.)

We need to imagine measuring the spin of a particle along three axes; x, y and z. Hold your hand up and make a shape that looks like the one in the following picture. You might remember it from science classes; it was used to help you understand Fleming's left-hand rule. For our purposes it does not matter which hand you use; it is just the shape that matters. I am going to use my left hand for sentimental reasons.

Now, imagine the palm of your hand is the measuring apparatus: your index finger the x axis, your middle finger, y and your thumb, z. At any moment you can move your hand to point in any direction and take a measurement. We will have to round up or down. Quantum mechanics is named 'quantum' because *all* the readings must be whole numbers. You will never see 10% spin in x, 90% in y and 85% in z; just ones, and zeros. The measurements for a Boson will always be 1,0,1 in some order. This is known as the '101' rule.

Now, we ask the question: does a particle have a definite spin before we take a measurement? The instinctive answer is yes, and this way of viewing things is known as realism. It seems obvious that even if we did not make a measurement, the particle would still have its spin; we just wouldn't know which type. Einstein explained realism by saying "I like

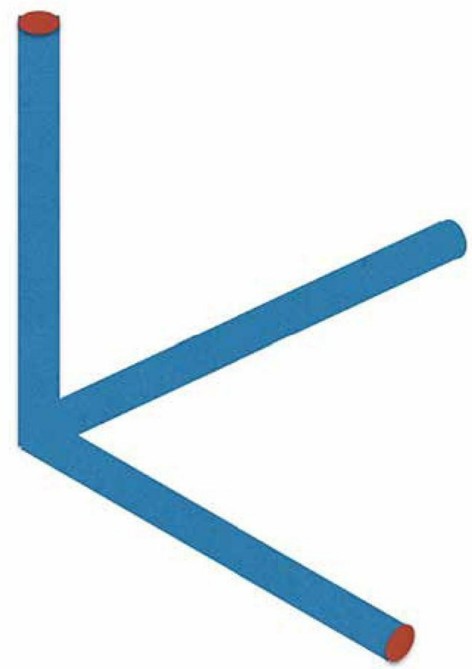


Kochen Specker

to think the Moon is there when I am not looking at it." But, how can we test his statement? How can we know something is there without taking a look? There is a way...

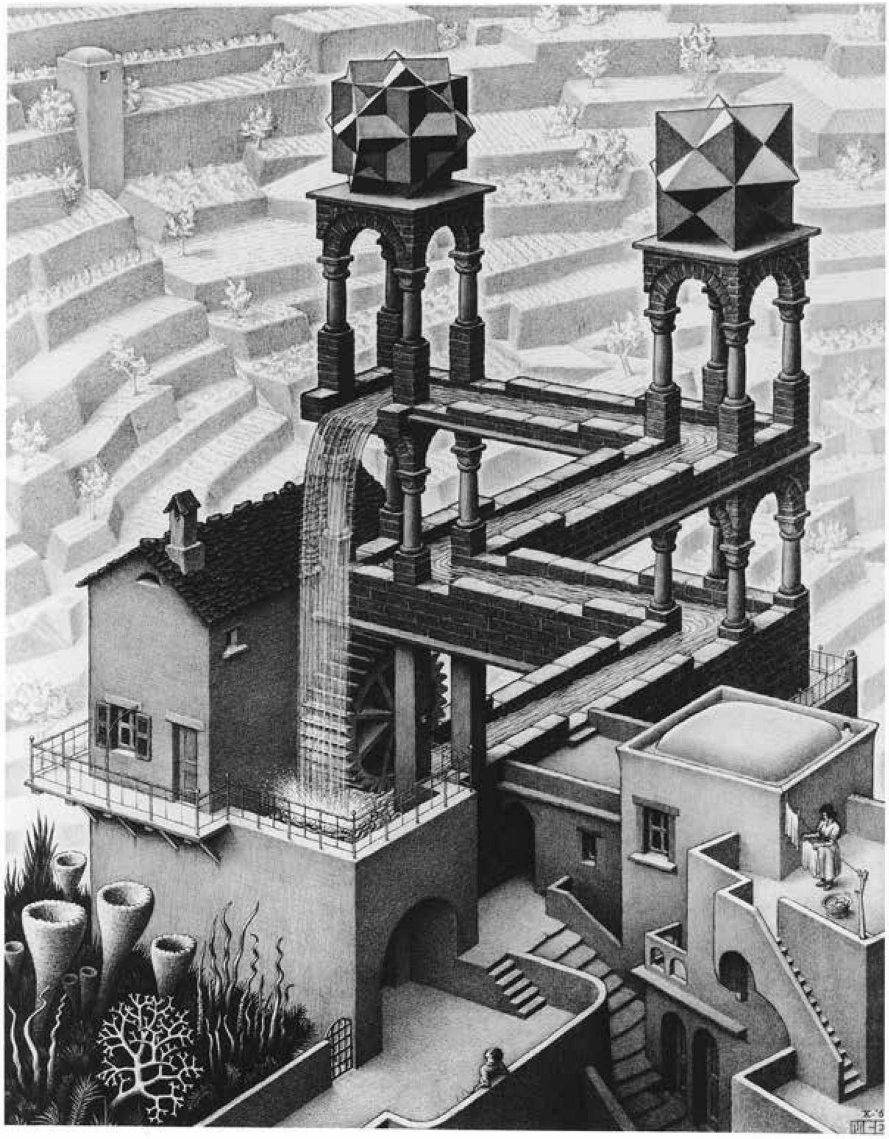
Let us suppose the particle had a definite spin before we measured it. Perhaps its spin points at the top left hand corner of the room. Imagine taking many measurements and seeing what happens. We can point our hand in any direction: top of the room, bottom left corner, bottom right corner and so on. Each time we point our hand in a direction we must get 1, 0 and 1 in some combination (110, 011, 101). The particles are 101 particles and this is an absolute rule.

Let's imagine doing the experiment. We fix the spin of a particle and begin to take measurements, noting the answers as we go. If we get a borderline condition we obey the 101 rule and give ourselves a 1, 0, 1 reading. As we move our hand to take measurements, a problem begins to emerge. Every now and again we obtain a measurement that conflicts. We chose a 1, 0, 1 when we were pointing our index finger towards the floor, but if we point the finger toward the door, we need that original middle number to have been a 1 for consistency. (The middle finger is now pointing in the direction the index finger pointed to for the first reading.) To fix the inconsistency we can change our original borderline decision to a 1,1,0. All is well and we continue. But, as we get over 30 measurements, we can't seem to find any way to make all the 1,0,1s fit together. After scratching our head for a while, we realize there might be no solution. And indeed there is not. This is the Kochen-Specker Paradox. The odd shaped cubes on the building in the Escher print are an example of one of these impossible figures.



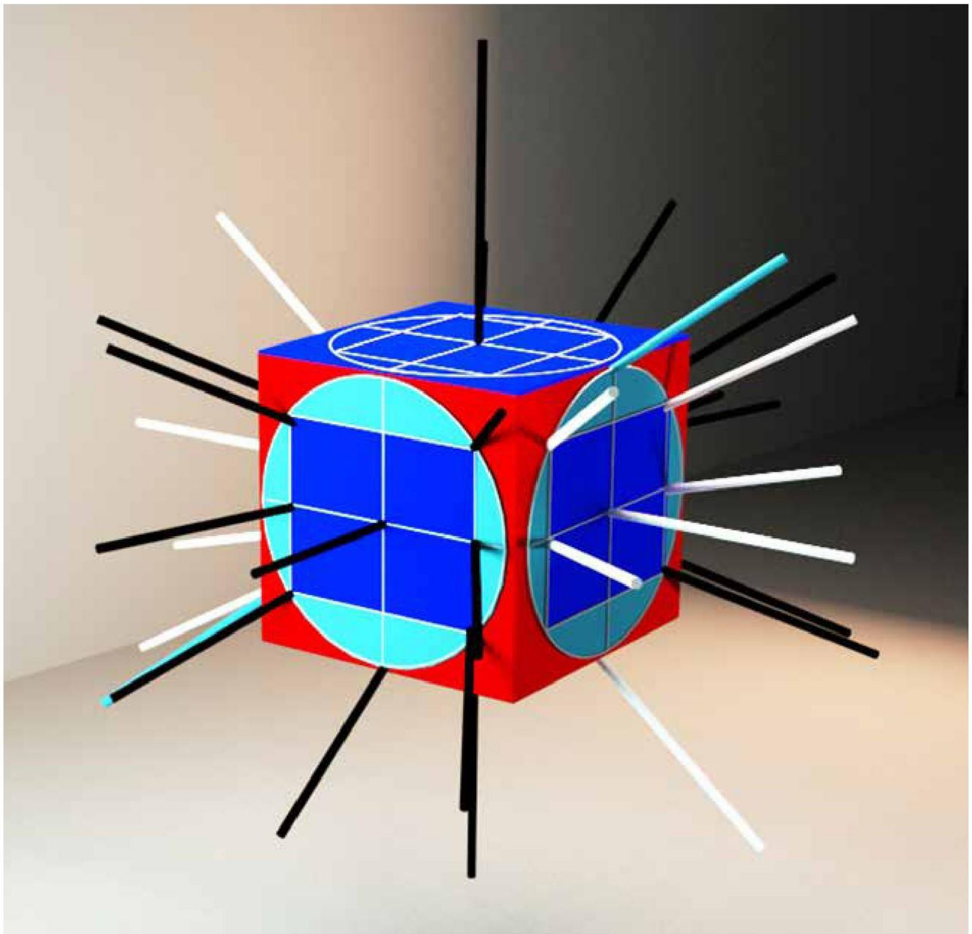
1 0 1 puzzle piece

An analogy to this problem is trying to solve a broken Rubik's Cube. There is a really mischievous trick you can play on someone: reverse two colors on a Rubik's Cube. You can easily do this by snapping one of the edge blocks out, turning it around and snapping it back in. When the colors are already muddled up this is not obvious. Now give your



M. C. Escher's Waterfall (Impossible Shapes)

friend the puzzle and they will spend hours trying to solve it! It can't be done because the puzzle is put together wrong. And in this matter nature is also put together wrong! With as few as 33 measurements it is impossible to construct a consistent three-dimensional shape that has 1s and 0s obeying the 101 rule in every place. The only way to complete such a shape is with a measurement that is both simultaneously zero and one: a paradox. We know what happens when we generate paradoxes. It means one of the original assumptions is false and, in this instance, the falsehood is that a particle has a definite spin before we measure it. It



Kochen-Specker Cube

cannot. It must make up its mind on the fly. Einstein would be horrified. Realism is violated by the quantum world: reality and measurement are intertwined.

The Kochen-Specker paradox shows us that a particle only makes its choice at the point of measurement. This does not prove it has free will as it might still be told what to do by some external entity. It's rather like the famous game show, *Who Wants to be a Millionaire?* The particle could answer the spin question in four possible ways. First, it could *know* the answer, but we have just proven it does not. Second, it could *phone a friend* obtaining the answer from some cosmic arbiter. Third, it could ask the audience and take a vote from all the particles around it. Finally, it could freely choose, without recourse to any of the other possible options – in other words, it would guess!

A guess would mean particles have free will; no extraneous influence or piece of information either on their person or from some external source could have any effect. We are now going to prove the particle *does* guess.

The Proof

Conway and Kochen construct their proof from a small set of axioms, which form a rhyme. The axioms are; twin, fin and spin.

If two particles are separated by a distance (fin) and entangled (twin), the spins of the particles (spin) cannot be determined by any information in *their* history of the Universe up to that point. The proof relies on a thought experiment.

Consider twin particles separated by a long distance. Physicists call this 'space like separation'. All this means is one particle is measured on, say, Earth and the other on Mars, so relativity is significant in the experiment. This may be impractical today but there is no reason the experiment could not be done in principle. In the future, our children could set up on the UN Moon base and fire one photon to a detector on Hubble II and the other to the future Mars Orbital Station. Farfetched? If you had told Einstein back in 1947 that in less than 70 years we would be able to measure individual photons by sending them down spun glass fibers to locations separated by 50 kilometers, involving a multidisciplinary team composed of American, German, French and Russian scientists, all working in harmony, he might have been equally incredulous.

As the proof introduces relativity we also need two imaginary rocket ships. They must be traveling below the speed of light, so no Star Trek Enterprise or Millennium Falcon. We will have to stick with an old-school spaceship, the Sulaco from *Aliens* should do the trick. They must travel in opposite directions, passing our Moon Base just as the scientists run the experiment.

Special Relativity shows our Universe has a strange property: there is no such thing as a simultaneous event for two observers – at least if they are separated by any distance. From the point of view of the first spaceship, the measurement on Mars occurs first. But from the vantage point of the second observer, the measurement on Hubble occurs first.

Now comes the proof by counter example. Let us suppose the particles *were* influenced by an outside effect and had no free will.

Let us say the Mars particle chooses its answer because of an external influence. Its Hubble twin must choose the same answer. There is no problem with this because the Hubble particle could have made its decision before the Mars particle, so the decision was not predetermined.

But in another frame of reference the choice is made in the opposite sequence. The Hubble particle chooses after the Mars particle. This is predetermination and it breaks the Kochen-Specker theorem.

You can reverse the whole analysis and see the same problem from the other point of view. There is a paradox here however you look at it.

The only solution to the paradox is that both particles make their choice without any information from an outside source; particles have free will. This means at least one new piece of information spontaneously appears in the Universe – a 'bit' of free will, so to speak.

You might think there is a problem because the first particle affects its twin, even if the second did not receive any outside influence. This would result in the Kochen-Specker paradox reemerging. There is a neat way out of this; time has no meaning for the particles. Or, I should say, relative time has no meaning and, therefore, has no effect. There is no concept of before or after between the particles. They live in a little bubble of space-time where the order of events has no meaning. The particles make their free choice together within this safe bubble, and the paradox is avoided. When we come to measure them, we see they both made a random decision together, but if we ask which made it first, the question has no meaning. There is no clock valid for both particles, so there is no possible answer to the question.

Conway and Kochen have proven sub-atomic particles have free will – or at least entangled bosons do. At this point, their argument becomes a philosophical one. They propose that these particles pass on this free will to larger entities in the Universe and ultimately to us. Although particles are small and insignificant, they are the fundamental building blocks of nature, and the butterfly effect multiplies up tiny variations in the microscopic world into the macroscopic events we see.

Although their theorem is very elegant, we still have to address the question of whether the experimenter has the true freedom to run the experiment in the first place: the determined determinist argument.

Russian Dolls

"Great fleas have little fleas upon
their backs to bite 'em.

And little fleas have lesser fleas,
and so ad infinitum.

And the great fleas themselves, in
turn, have greater fleas to go on.

While these again have greater
still, and greater still, and so on."

Augustus De Morgan

Free Will Universe

I believe we live in a Universe where information comes into existence through the creative endeavors of human beings. When Andrew Wiles discovered his solution to Fermat's Last Theorem, he did something a computer cannot do and demonstrated non-computational thought. But there is an alternative explanation put forward by the determined determinists.

Daniel Dennett – the standard bearer for this camp – believes everything in the Universe is entirely determined. He argues there is no place in the laws of nature for free will to arise.

Both sides of the argument agree Turing prohibits a general-purpose machine from solving all mathematical problems, but that seems to be the extent of agreement. The determinists solve the Wiles Paradox by arguing he is a special purpose machine, perfectly able to find answers to non-computable problems. The Turing prohibition only applies to general purpose machines. Let us run a thought experiment to see what sort of Universe we would live in if special purpose machines were the answer to this puzzle.

If the Universe is determined, it can be modeled as a single algorithm. If everything in the Universe evolves according to a set of rules, it will run like a giant piece of clockwork or one large computer game. Each solar system, planet, and individual mathematician would evolve along preordained lines. Mathematicians would operate as software subroutine and would rely on further subroutines to explain the beating of their hearts and the way the molecules of their body interact.

If our Universe were organized in this way:

This Universe *could not* discover solutions to arbitrary problems.

This Universe *could* be preprogrammed with every theory we could ever discover within it. (There would be no arbitrary problems.)

This argument neatly sidesteps Turing's theorem by specifying there is no such thing as an arbitrary problem – a random problem picked from the infinite set of problems. At the same time, it sets certain characteristics of such a Universe and I believe we can test these...

A computable Universe must already know the solution to every problem it will encounter above the logic limit: It cannot discover knowledge on the fly. For many problems, a small number of fundamental rules can account for everything. Although our galaxy and the beautiful nebulae we see through our telescopes look complex, they might be the result of some such simple set of rules – just like a fractal. That's Stephen Wolfram's solution to the mystery of our Universe. But some problems are complex. The solution to Fermat's Last Theorem is an 80 page document consisting of 5 million bits of information. All this must be stored somewhere in the Universe. It might not be stored as a string of bytes, it could be found in a set of equations governing the motion of the atoms such that at some point – in 1995 to be exact – they all line up in Andrew Wiles' brain to direct his fingers to type out the proof. In this case, the Universe has solved a mathematical puzzle because it was specifically set up to do so from the time of the Big Bang, but this raises three questions:

Where does the Universe store this enormous amount of information?

How does The Universe hold the information reliably?

How did the pre-Universe solve the problem, so it might program the Universe at the moment of the Big Bang?

The first question is probably answerable. The Universe is a big place and could store sufficient information to solve the mysteries that puzzle the inquisitive creatures that inhabit its planes. There are many practical problems to consider, such as how to preserve the information through all the strange evolutions of our Universe; inflation, star formation, and so on. But it could be done.

The second question is insurmountable and presents the counter argument to the determinists. Our Universe appears to be composed of non-deterministic objects. Such objects exist in the mathematical world; Kochen-Specker cubes, for example. Unfortunately for the determined determinist, bosons behave according to the same principles. In case you're thinking thinking bosons are rare, light is formed of bosons. Our whole existence is surrounded by non-deterministic physics. Therefore, your actions are not predetermined by anything in your local corner of the Universe – the past light cone if you want to be strict about the physics.

The determined determinists are a determined bunch. Just because the information that determines your actions cannot be encoded by the particles you are made from, does not mean you are free. The information could be stored in parts of the Universe we cannot see, or held outside the Universe in some sort of cosmic hard drive. Every creative event in the Universe would be specified in this store.

But this begs the third question: How was this store of information generated in the first place? If a Universe contains creative things – as our Universe does – there is no way to computably generate the necessary determinist store of information. The Universe has free will because there is no deterministic process that could generate it.

If our Universe were a Turing machine, everything within it would be too. Think about the deterministic clockwork argument I gave earlier. If you try to construct a better – say a more random – machine inside a Turing machine, an observer could simply ignore the better machine hidden within it, and watch the outer machine work. The outer machine will predict the operation of the inner machine perfectly, even if the inner machine is fiendishly complicated. We have to consider the machine on which our human software runs. Our bodies, our minds, all that we are, is software running on the Universe's hardware of quarks and photons. If the hardware is deterministic, then so is our software. And if the hardware is deterministic, there can be no creativity within the Universe.

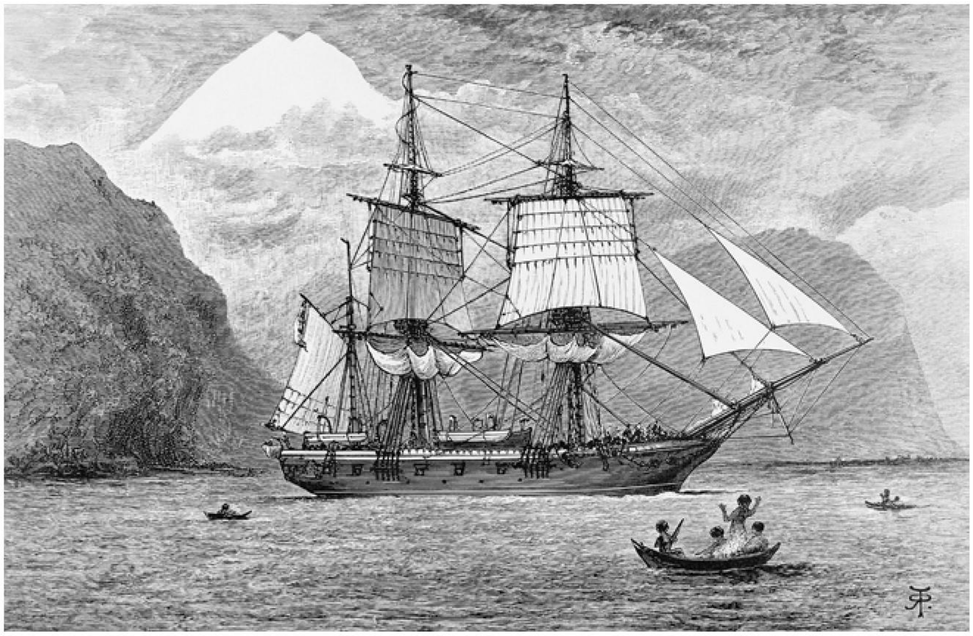
So the free will camp has an argument easily as frustrating as the one deployed by the determinists. Every time a determinist asks, "How do you know you were not always going to do that?" the free will believer can reply, "You asked me a question. If this dialogue is to have any significance, then we must exist in a rational Universe and, therefore, the laws of information give us creativity and free will. If you believe we are fully determined, there is no point in my answering your question."

I reason. Therefore, I have free will.

The Universe is not a machine.

Chapter 16

THE QUEST FOR KNOWLEDGE



Darwin's Beagle

"Sometimes I've believed as many as six impossible things before breakfast."

Queen of Hearts in

Lewis Carroll's Alice

"It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change."

Charles Darwin

We celebrate creativity with many competitions and prizes. I have been a student of problem lists of over the years. Here is my list of the problems remaining open in the modern world. I've tied it in with other lists where relevant, and indicate what you might win if you were to solve one. As I was writing this book, a few of the questions were answered; the Higgs Boson was discovered and the Poincaré Conjecture proven. I will keep the list up to date on the web site.

1. **Mathematics** [Fields Medal]
    - 1.1 The Birch and Swinnerton-Dyer Conjecture.[C]
    - 1.2 Hodge Conjecture [C]
    - 1.3 Navier-Stokes Equations [C]
    - 1.4 A proof or disproof of P =NP [C]
    - 1.5 The Poincaré Conjecture [C] – Solved
    - 1.6 Riemann Hypothesis [h8 C]
    - 1.7 Yang-Mills Theory [C]
    - 1.8 Can we understand and solve all 23 Hilbert Problems [h1-23]
    - 1.9 Goldbach Conjecture [h8]
    - 1.10 Is mathematics fundamental to or simply a good model of our Universe? [H6]
    - 1.11 Is mathematics an emergent property in our Universe or causal? Which is more fundamental?
    - 1.12 Fermat's own original proof of his theorem!

2. **Physics** [Nobel Prize for Physics]
    - 2.1 Are there many worlds or just one?
    - 2.2 Does quantum collapse have meaning?
    - 2.3 Will we find the Higgs-Boson? (Provisionally yes, 2012)
    - 2.4 Do we need quantum gravity to explain human thought?
    - 2.5 Will we observe gravitational waves, and what is the current explanation for gravitational noise?
    - 2.6 What causes the arrow of time and the asymmetry of physical laws?
    - 2.7 Do the constants of physics change over time?
    - 2.8 Is quantum computation sufficient to simulate the universe, or is the universe non-computational?
    - 2.9 Can magnetic monopoles exist?
    - 2.10 What is meant by quantum non-locality, a.k.a., spooky action at a distance?

- 2.11 Is there a theory that would unite gravity with the other three forces: a Theory of Everything?
- 2.12 Is Schrödinger's cat alive or dead in the box?
- 2.13 Does ball lightning exist and can it be made in the laboratory?

3. Cosmology Nobel Prize for Physics

- 3.1 What is the nature of Dark Matter?
- 3.2 What is the nature of Dark Energy?
- 3.3 What is Dark Flow?
- 3.4 The slingshot anomaly.
- 3.5 Did inflation really happen?
- 3.6 Was there a singularity at the origin of our Universe, and what happened before the first second? An eternity?
- 3.7 Can any information travel faster than the speed of light?
- 3.8 What is the cause of the Pioneer anomaly? (solved in 2011)
- 3.9 Are there aliens?
- 3.10 The Goldilocks question. Why are the cosmological constants so finely tuned?
- 3.11 Do real numbers exist or is our Universe quantized?
- 3.12 Why is there little antimatter?
- 3.13 What are cosmic rays and where do they come from?
- 3.14 What was the WOW signal?
- 3.15 Does the fine structure constant vary over time?
- 3.16 If we live in an infinite Universe, why don't we see more strange things?
- 3.17 Why is the cosmic background radiation so smooth?
- 3.18 How can we explain the lack of total smoothness of the cosmic background radiation!
- 3.19 Is there an explanation for any detail in the cosmic background radiation map?

4. Engineering Nobel Prize for Chemistry, Turing Award

- 4.1 Can we achieve economic nuclear fusion?
- 4.2 Will we realize cold fusion? (Partially demonstrated)
- 4.3 Can an amateur get to the moon?
- 4.4 Can an amateur collect a rock from the moon? X
- 4.5 Will we make an Artificial Intelligence?
- 4.6 Can we make a Tricorder? X
- 4.7 Can we power the world from renewable sources?
- 4.8 Will robots go to war in the future?

4.9 Can we make a robot indistinguishable from a human and cross the uncanny valley?
4.10 What is the tallest building we could build on planet Earth?
4.11 Will we routinely use flying cars by the end of this century?
4.12 Will we have a base on the Moon or Mars in this century?

5. Biology Nobel Prize for Medicine

5.1 Why is the placebo effect so strong?
5.2 Can we cure the common cold?
5.3 Is there a generally effective treatment for cancer?
5.4 Can we find a vaccine against HIV?
5.5 Can we cure endemic diseases such as malaria, or is it an arms race?
5.6 How plastic are our genes and is epigenetics a significant factor?
5.7 Can we cure dementia? L
5.8 Can we make a desktop gene sequencer? X
5.9 Will we prove the Kurzweil Hypothesis that technology will allow us to live forever?
5.10 How old will we live to with a reasonable quality of life?
5.11 Can genes jump between organisms? Even participating in whole scale fusion?
5.12 Will we be able to grow organs?
5.13 Will we clone a human from an adult?
5.14 Can we clone a dinosaur?

6. The Mind Nobel Prize for Medicine

6.1 Does the Flynn Effect mean we are really becoming more intelligent?
6.2 What is the nature of consciousness?
6.3 Do we have free will?
6.4 Which is more important: Nature or Nurture?
6.5 What is humor for?
6.6 Do some people have photographic memory? (yes, recent)
6.7 What is the purpose of sleep and, in particular, dreams?
6.8 What is understanding?
6.9 Do we ever truly know something?
6.10 How does the brain think?
6.11 Is the brain a quantum device?
6.12 Why do we get stressed?
6.13 Why are some people more intelligent than others?

6.14 What limits our ability to concentrate and work hard mentally?

7. The Ancient World Pulitzer prize for History

7.1 What is the Linear-a script discovered in Crete?
7.2 Where are the ruins of the Light House at Alexandria and indeed Alexandria itself?
7.3 What is the location of the Lost City of Atlantis if it is not a myth?
7.4 Will we ever find King John's Treasure?
7.5 What is the truth to the legend of El Dorado?
7.6 Why were the pyramids built?
7.7 What was the purpose of Stonehenge and who built it?
7.8 How many books and how much knowledge have we lost?
7.9 Did King Arthur and Camelot exist in any real way?
7.10 Why did the people of Easter Island build their statues?
7.11 Was there an ancient flood, suggested by the Bible and other ancient texts?
7.12 Are the Seven Wonders of the Ancient World lost forever?

8. The Modern World Nobel Peace Prize or Prize for Economics

8.1 Is there a best political organization for a country?
8.2 What is the best political balance of federation and autonomy?
8.3 The Black Swan Effect: Why do improbable things happen?
8.4 Is there a right way to run the economies of the world?
8.5 When is it right to intervene in a conflict, and when is it best to leave a country to its own devices?
8.6 What is the best way to choose a political representative?
8.7 Will we ever abolish war?
8.8 Why is the gap between rich and poor increasing in most of the world today?
8.9 How powerful should states be compared with world organizations?
8.10 Is there a right level of tax?
8.11 Are morals absolute or relative: euthanasia, abortion, gay marriage, eating meat?
8.12 What will we do about our aging population?

9. Planet Earth Goldman Prize for the Environment

9.1 Is man-made global warming real, and if already proven will we ever persuade the US government?
9.2 What caused the Tunguska Explosion?
9.3 What caused the extinction of the dinosaurs?
9.4 Can we predict earthquakes or eruptions?
9.5 What caused the reversing of the poles and when will the next one occur?
9.6 What was the origin of life on Earth?
9.7 Will we be wiped out by an asteroid before we build a suitable defense?
9.8 Can we grow enough food to feed the planet? L
9.9 Can we give clean water to everyone on the planet? L
9.10 Is increasing air travel compatible with survival of the planet? L

10. Philosophy Nobel Prize in Literature and others

10.1 Is there a God?
10.2 Where did we come from if we are not made by a god? And if we were, then where did God come from?
10.3 Where do morals come from?
10.4 Is there a reality?
10.5 Is there life after death?
10.6 Do we have free will?
10.7 Is beauty in the eye of the beholder?
10.8 What is the meaning of life, the Universe and everything, other than 42?

11. Conspiracy and Paranormal

11.1 Is there anything going on in the Bermuda Triangle?
11.2 Do aliens make crop circles? (disproven hoax)
11.3 Who was Jack the Ripper?
11.4 Can the mind bend spoons? (hoax, admitted)
11.5 Does the government suppress UFO existence?
11.6 Why was the Mary Celeste abandoned?
11.7 Is the Turin Shroud that of Christ?
11.8 Do the Abominable Snowman and Sasquatch exist?
11.9 Are there ghosts?
11.10 Is there a paranormal?
11.11 Do aliens live amongst us?
11.12 Was there a conspiracy in the shooting of JFK?

Cross reference to other lists

Hn: Hilbert's Problem
C: Clay Mathematics Millennium Prizes
X: XPRIZE
L: Longitude Prize

Nobel Prize Medal

> *"If I could explain it to the average person, I wouldn't have been worth the Nobel Prize."*
>
> Richard P. Feynman

Awards for Discovery

People like prizes. Competition drives humans forward in a way we don't properly understand. In film, we have the Academy Awards, whilst on the web we have The Webbies. Some prizes, such as the Nobel Prize, Fields Medal and Pulitzer Prizes, have a long and distinguished history, while others such as the XPRIZE are more recent creations. Some prizes, such as the Ig Nobel Prize and the Golden Pineapples were mainly created for their humorous value. Prizes are not a recent phenomenon. The Longitude Prize, originally won by John Harrison, is being revived in Britain in 2014 to mark its 300th anniversary. The original prize, £10,000 in its day, was awarded by the British government for making a device that allowed ships to determine their East-West position (a sextant only gives north-south). The 2014 prize is £10m pounds and the topic will be chosen by public vote! Here is a small history of some of the more famous prizes.

Nobel Prizes

Alfred Nobel spent his life developing weapons and explosives. His laboratory was built in the middle of a lake with a bridge running to it, so if he blew himself up doing an experiment, only he would die. He managed to stabilize nitroglycerine by mixing it with saltpeter and created

Pulitzer Medal

dynamite. This was used in the mining industry but also extensively in weaponry, so he came to be known as the Merchant of Death during his lifetime.

To be known as the merchant of death would have a profound effect on anyone. As Nobel pondered the balance of his life's work he decided to do something positive with the huge wealth he had accumulated. On his death in 1896, he willed his entire fortune to create the awards we now call Nobel Prizes.

There were five original prizes; physics, chemistry, peace, physiology or medicine, and literature. A newer economic science prize is awarded by the Royal Swedish Academy of Sciences.

You must be alive to receive a Nobel Prize; a few have been awarded posthumously because the laureate died after the winner was announced but before the award ceremony. The work must have been proven experimentally, and although originally it was supposed to be for discoveries in the previous year, nowadays a theory must have stood the test of time. Consequently, winners tend to be quite old. The prize must be for something with practical applicability – Einstein received his Nobel Prize for the Photo Electric Effect, rather than his more famous Theory of Relativity. The judges evidently thought particles more practical than planets! The prize is usually awarded to a maximum of three people. This has produced some controversial results but despite this the Nobel Prize is the uncontested top prize in science.

Pulitzer Prize

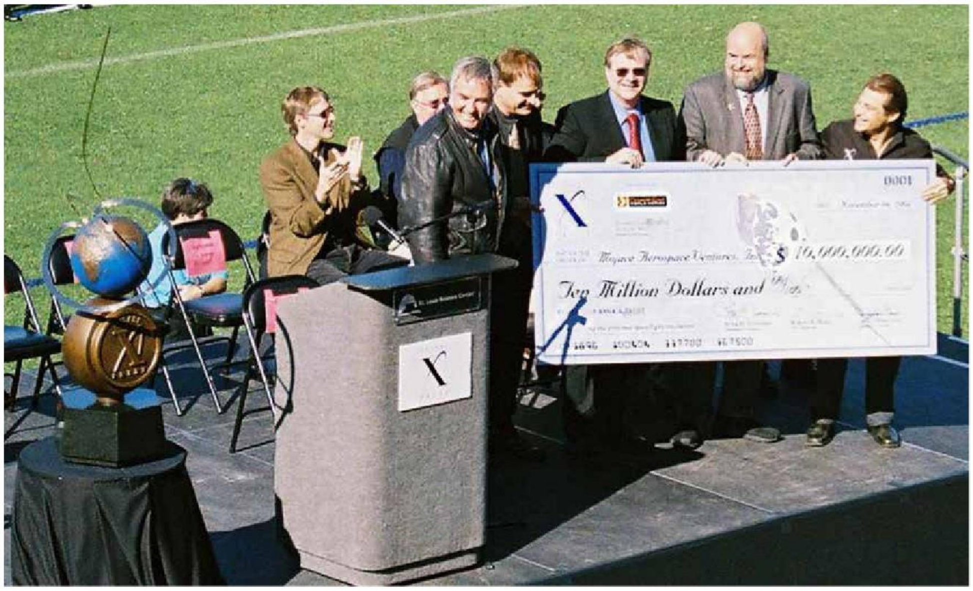
A Pulitzer Prizes is to the arts what a Nobel Prize is to science. Again the Prize was the result of a bequest. They are awarded in the fields of music, art and literature. Unlike Nobel Prizes, where there are no public nominations and you might wait a lifetime for the phone call, you enter your name for a Pulitzer Prize. Most people associate the term Pulitzer Prize winner with journalism, but about 25 Pulitzers are awarded each year. You must be a US citizen to enter.

Turing Award

Originally set up by the Association of Computer Machinery, this award comes with prize money of $250,000, supported by Google and Intel, and goes to a person who significantly advanced computer science or artificial intelligence in the previous year. It is considered the Nobel Prize for computing.

XPRIZE

XPRIZEs are awarded for technology and bear the democratic stamp of the Internet age. Anyone can propose a challenge but they must also provide the prize money! It's big money. The Ansari XPRIZE for the first



XPRIZE First Award Ceremony

non-governmental organization to put a man in space was $10 million, awarded in 2004. There are a growing number of XPRIZEs, including, at the time of writing:

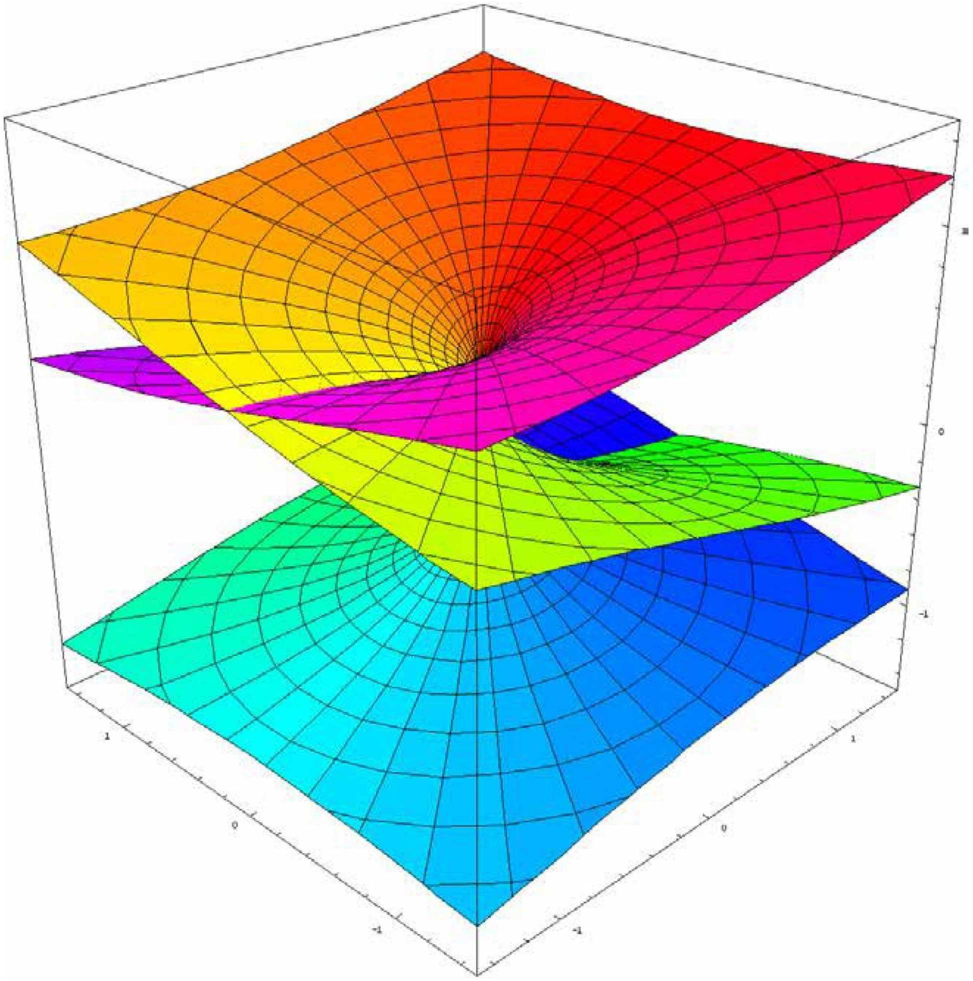
- Google Lunar XPRIZE, $30m to put a rover on the Moon.
- Qualcomm Tricorder XPRIZE, $10m to make your mobile phone into a hand held medical health scanner, similar to the Star Trek tricorder.
- Nokia Sensing CHALLENGE, $2.25m to build a hand-held medical scanner.
- Wendy Schmidt Ocean Health XPRIZE, $2m to create a method to measure the ocean's pH.

Fields Medal

The equivalent of a Nobel Prize for mathematics is a Fields Medal. Joseph Field provided the money and helped set up the prize. Today it is administered as part of the International Mathematical Union. You must be under 40 to receive the prize. Andrew Wiles was 45 when he solved Fermat's Last Theorem, so they created a special prize for him called a Fields Fellowship. Until recently only men had received the prize. However in 2014 Maryam Mirzakhani won the prize for her work on the geometry of Riemann surfaces.

TRANSIRE SVVM PECTVS MVNDOQVE POTIRI
ΑΡΧΙΜΗΔΟΥΣ

Fields Medal



Riemann Surface

Chapter 17

THE FUTURE

Omar Khayyám

“Prediction is very hard, especially about the future.”

Niels Bohr

“The Moving Finger writes: and, having writ, Moves on: nor all thy Piety nor Wit Shall lure it back to cancel half a Line, Nor all thy Tears wash out a Word of it.”

Rubaiyat of Omar Khayyam,
Edward FitzGerald

رباعیات عمر خیام

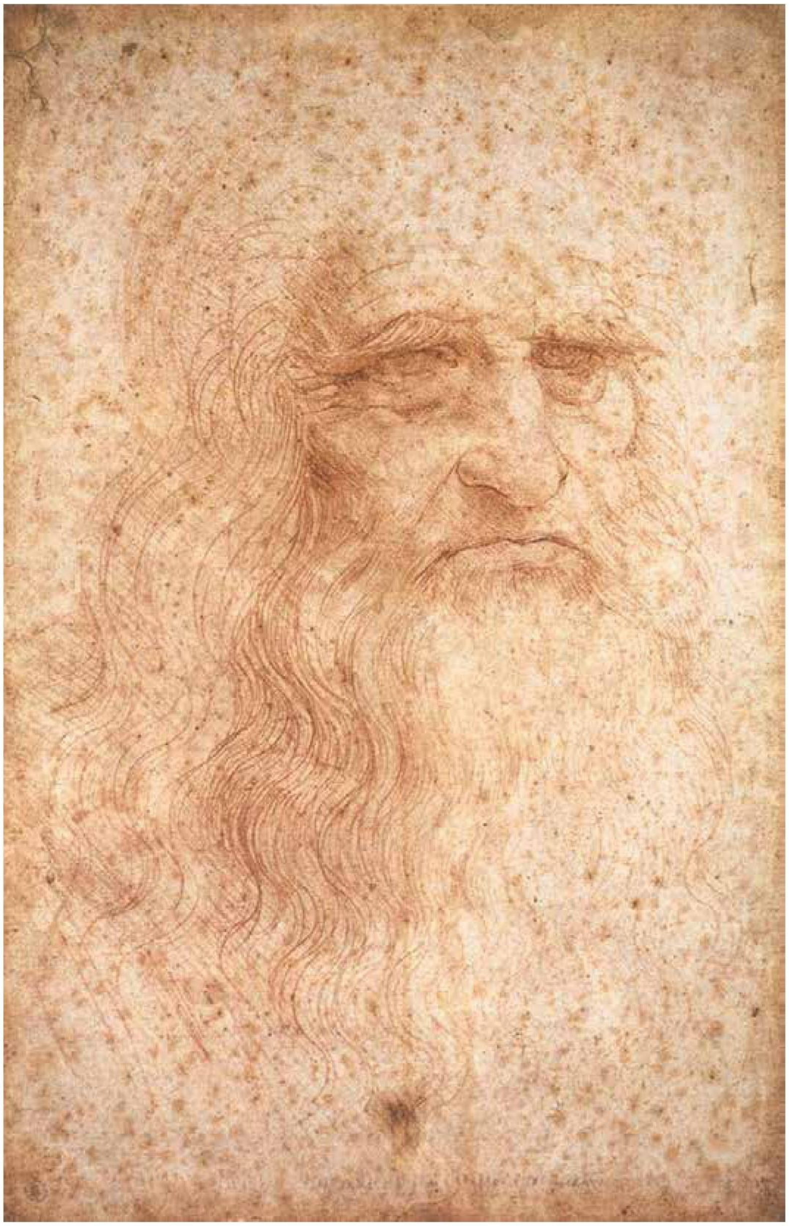
I remember when I was eight years old, being asked to draw a vision of the world in the year 2000. In my the home of the future, rather than going to the shops to get milk, orange juice and cornflakes, they would arrive by pipe. These days I know about microbiology and realize this would have been highly impractical and perhaps rather dangerous. I could claim some premonition of the Internet at this point; no self-respecting science book is complete without one of these!

Of course, the truth is I had no more idea of the way things would turn out than anyone else. Now that I am a little older let's see how much trouble I can get into predicting the future.

I think we will build thinking machines – AIs – using our insights into the operation of the brain. They will not be like the computers of today but will still be physical devices. There is nothing overtly spiritual in my conception of the way we operate, but I am arguing that the human mechanism is more complex than a digital computer. Building these machines will be hard, and they will not be 'machines' in the sense I have used throughout this book. They will be minds.

When we build AIs that think and feel, will they acquire 'human' rights? Might one of my grandchildren fall in love with an AI, perhaps even marry one? On the darker side, how will they view us: what place would we have in their world once we had brought them into being? However, I think this process of building an AI will be hard and in one hundred years' time we will still be struggling with the problem.

In this book, I have presented a way to understand the creative process within our Universe. It relies on the existence of non-computable processes in our brain and in the physical laws which govern them. Currently, the laws contain a big hole. Although we can, perhaps, see where freedom might come from – through randomness and non-determinism – we don't understand where the will emanates to shape the Universe. Over the next thirty years, I think we will begin to understand this and see how creativity relates to the Universe we observe. I am not suggesting any anthropic principle, or some grand interaction between mankind and the Universe, just an important simple freedom: That we humans are free to think and do as we please. When I choose to lift my arm and raise a glass of wine with friends, this is my choice. I am the cause. The effect is the displacement of my arm, causing photons and gravitational waves to ripple out across the Universe, and in that sense I freely affect my environment.



Da Vinci, Self Portrait

> *"A good painter is to paint two*
> *main things, men and the*
> *working of man's mind."*
>
> Leonardo da Vinci

Appendix 1

Acknowledgments

Front Matter

Cover	Vladislav Ociacia
Spine Equations	Illustration by Arabella Tagg
Author Photograph	Arabella Tagg
ACPMM, Wolfson College Cambridge	James Tagg Personal Collection, Course changed name to ACDMM in 1990.
Mathematical Bridge, Cambridge	Hipgnosis, www.shutterstock.com
Introductory Image	Photograph by James Tagg

Chapter 1

Computer versus Human	Blutgruppe/Corbis
Kasparov versus Deep Blue	Louie Psihoyos/Corbis
The Music of Emily Howell	Kind permission of David Cope and Centaur Records. Emily Howell: From Darkness, Light. Picture and Audio Clip
IBM's Watson Plays Jeopardy	Associated Press Carol Kaelson/ Jeopardy Productions, Inc.
Watson Questions and Answers	Illustration by James Tagg
Steve Wozniak	TIM CHONG/Reuters/Corbis
Turning Images to Music	Credited to: Maxim Dupliy, Amir Amedi and Shelly Levy-Tzedek
Brain Image of Fish Hunting Prey	This work (or this video) was published from Kawakami lab in National Institute of Genetics , Japan (Muto, A. et al. Current Biology 23, 307–311, 2013)".
Babbage Difference Engine No. 2	Photograph by James Tagg @ The Computer History Museum
19th Century Calculators	Wikimedia, Ezrdr, CC3
Model of the Antikythera Mechanism	Wikimedia, Geni, CC3

Moore's Law	Credited to Ray Kurzweil, CC1
3D Chip	Kind permission Intel Press Department
Richard Branson	kathclick, www.bigstock.com
ELIZA, DOCTOR	Opensource project encapsulated into widget by James Tagg
IQ Test	Illustrated by James Tagg based on a Wechsler example question
Metal Puzzle	www.shutterstock.com, fdpress
Hole in the Wall Experiment	Courtesy Philippe Tarbouriech/Hole-in-the-Wall Education Ltd.
One Laptop per Child	One Laptop per Child project
Piano Practice	Photograph by James Tagg
Dan McLaughlin	Kind permission of Dan McLaughlin, www.thedanplan.com
Astrological Clock, Hampton Court	Wazzaman, Wikimedia, CC3
Lava Lamp	Sean Gladwell, www.shutterstock.com
Steve Jobs Collage	Kind permission: www.village9991.it
"Ascending and Descending"	© 2014 The M.C. Escher Company-The Netherlands. All rights reserved. www.mcescher.com

Chapter 2

Afghanistan Stability/COIN Dynamics	US Government, Joint Chiefs of Staff, PD
McChrystal in Kabul	USA Navy Photo, PD
Gettysburg Address as PowerPoint	Kind permission of Peter Norvig
Space Shuttle Columbia Crew	Credit NASA
Shuttle Tile	US Government, PD
Shuttle Images	Credit NASA
Searle's Chinese Room	Illustrated by James Tagg
Black Box Diagrams	Illustrated by James Tagg
The Miracle Worker, Helen Keller	Associated Press
Human Person, or is it?	Associated Press
New Yorker Dog Internet Cartoon	New Yorker © Condé Nast Licensing

Chapter 3

Body Language	Kind permission of Conference on Communication and Body Language.
Ronald Reagan and Mikael Gorbachev	US Government, PD
Höfði House in Reykjavik	Wikimedia
Fake or Real Smile	Bigedhar, www.bigstock.com
Yasser Arafat and Shimon Pérez	UPI
Learning Swedish, The Two Ronnies	Kind permission of BBC, hosted on YouTube
Scripts of the World	Illustrated by James Tagg
Chinese Traditional and Simplified	Illustrated by James Tagg
Great Comedy Videos	Kind permission BBC hosted on YouTube

Chapter 4

Child Having EEG	dblight, www.iStockphoto.com
X-Ray of Rontgen's Wife's Hand	Wikimedia
Lego Cubes Under Ultraviolet Light	www.public-domain-image.com
Pit Viper	abcphotosystem, www.shutterstock.com
Einstein's Brain	Wikimedia
Thermal Image of a House	Fotoflash, www.bigstock.com
Flowers in Ultraviolet Light	Bjørn Rørslett
Functional MRI, Response	National Institute of Mental Health, Wikimedia, PD
Functional MRI: Working Memory	Kind permission John Graner, Neuroimaging Department, National Intrepid Center of Excellence, Walter Reed National Military Medical Center, 8901 Wisconsin Avenue, Bethesda, MD 20889, USA
McGill Diffusion Tensor Image	Thomas Schultz, Wikimedia, CC3
Functional PET	Jens Maus, Wikimedia, PD
Organization of Your Brain	www.shutterstock.com and James Tagg
Visual Processing System	Illustrated by James Tagg includes www.shutterstock.com components.
Impressionist Painting, Monet Haystack	Wikimedia, PD
Frogs Eyes are Very Sensitive	Michiel de Wit, www.shutterstock.com
Color is Not an Absolute Sense	Illustrated by James Tagg
McGurk Effect	Kind permission of BBC, hosted on YouTube
Penrose Steps	James Tagg, Sketchup Model
Scintillating Blobs	Illustrated by James Tagg
Selective Attention video link	Kind permission to link provided by Daniel Simons. DVDs can be purchased from www.viscog.com
Tiger Woods Swing video	Kind permission of www.craighansongolf.com
Neural Network	Illustrated by James Tagg
Synapse	Meletver, www.bigstock.com
Paramecium	micro_photo, istock
Quantum Tubulin	Kind permission Travis Craddock
Tubulin Molecule	Wikimedia

Chapter 5

Chimpanzee and Typewriters	chippix, www.shutterstock.com
There are Holes in the Sky poem	Spike Milligan Enterprises
Spike Milligan	TopFoto[ ]
Lewis Carroll's Jabberwocky	Lewis Carroll, Out of Copyright
Lewis Carroll's Jabberwocky	Lewis Carroll, Out of Copyright
Word's Verdict on the Jabberwocky	Illustrated by James Tagg

Loch Ness Monster Picture	Unknown, Hoax
The Loch Ness Monster's Song	From Glasgow to Saturn (Carcanet, 1973) also published in Collected Poems (Carcanet, 1990) Reprinted by permission of Carcanet Press.
Dyslexic Poem	Kind permission of the copyright holder: The Journal of Irreproducible Results, the science humor magazine, www.jir.com, 1994 and 2000, via the author Jerrold H. Zar
Starry Night, van Gogh	Wikimedia, PD
Game of Battleship	www.shutterstock.com
Jesse our Creative Kitten	James Tagg

Chapter 6

Orangutan and Kitten	Chris Butler, www.bigstock.com
Twin Guards	Manamana, www.shutterstock.com
Groucho Marx	Library of Congress, PD
Euclid's Elements, Oxyrhynchus Papyrus	Wikimedia, PD

Chapter 7

Mandelbrot Set	Steve Buckley, www.shutterstock.com
Bubble Sort Ballet, video	Created at Sapientia University, Tirgu Mures (Marosvásárhely), Romania. Directed by Kátai Zoltán and Tóth László. In cooperation with "Maros Művészegyüttes", Tirgu Mures (Marosvásárhely), Romania.
Maze	Vasilius, www.bigstock.com
Travelling Salesman Problem	Illustrated by James Tagg, map from www.bigstock.com
Rubik's Cube	Photograph by James Tagg
Complexity Scale	Illustrated by James Tagg
Butterfly Beginnings of a Tornado?	saichol chandee, www.shutterstock.com
Trajectories	Kind permission Steinn Sigurðsson (1991)
Poincaré Portrait	Eugène Pirou, Wikimedia, PD
Blue Marble, Weather Patterns	Reto Stöckli, NASA Earth Observatory
Lorenz Attractor	zentilia, www.bigstock.com
Nebula	Credit NASA
Cellular Automaton	Weisstein, Eric W. "Cellular Automaton." From MathWorld--A Wolfram Web Resource.
Conway's 'Life'	James Tagg screen capture of an MIT opensource project

Chapter 8

Hilbert's Hotel	Karamysh www.bigstock.com
Spears and Hunters	Munduruku, Wikimedia cc2.5

Traversing an Infinite plane with a Line	Illustrated by James Tagg
Spear and Hunter	Illustrated by James Tagg
Hilbert Hotel Video Link	Kind permission of BBC, hosted on YouTube
Holding Infinity in Your Hand	Photograph by James Tagg
Number Quiz 1	Illustrated by James Tagg
Number Quiz 2	Illustrated by James Tagg
Donate a Random Number	Illustrated by James Tagg
Which Number is Random	Illustrated by James Tagg
Smallpox Virus	3d4Medical.com/Corbis
Smallpox Child	Associated Press, SANTOSH BASAK

Chapter 9

Donald Rumsfeld	US Army, Wikimedia, PD
Kurt Gödel, any Likeness is Accidental	Unknown, Wikimedia, PD
IAF Rule 164	IAF rules excerpt
PM	Paul Hermans, Wikimedia, CC3
Amazon Listing for PM	Amazon excerpt (not in print version)
1+1 = 2, PM (1)	PM excerpt
1+1 = 2, PM (2)	PM excerpt
Konigsberg's Bridges	Bogdan Giuşcă, Wikimedia, PD
Peano Portrait	Materialscientist, Wikimedia, PD
Beer Mug, Table and Chair	Photograph by James Tagg
Einstein and Gödel	Kind permission of the Archive of the Institute of Advanced Study

Chapter 10

Alan Turing Portrait	National Portrait Gallery
Enigma Machine	Sperling, Wikimedia, PD
Can you decode this?	Illustrated by James Tagg
Correct the code	Illustrated by James Tagg
Lego Turing Machine	www.LegoTuringMachine.org
Old Fashioned Relay Mechanism	Wikimedia, Signalhead, CS3
3D Printing Machine	360b / www.shutterstock.com
Block Print from 'No Silver Bullet'	Wikimedia, PD

Chapter 11

Fred Brooks	Copyright owned by SD&M, Wikimedia CC3
Web Page (James Tagg's Home Page)	Screen capture by James Tagg
Dilbert Software Specification	DILBERT © 2006 Scott Adams. Used By permission of UNIVERSAL UCLICK. All rights reserved.
Long Multiplication	Illustrated by James Tagg
A Hypercube in Two Dimensions	Mouagip, Wikimedia, CC3
Impossible Shapes, Devil's Tuning Fork	Illustrated by James Tagg
Halting Program	Illustrated by James Tagg
Four Color Problem	chas zzz brown, Wikimedia, CC3
Dalek Trouble	Birkett 1981, Permission Punch
Augustus De Morgan Poem	Augustus De Morgan, (pd)

Word Puzzle	Gennadií Makanin Excerpt of paper on Word Puzzles
Creative Inoculation	Illustrated by James Tagg
Jackson Pollock	Albright-Knox Art Gallery/CORBIS, Pollock-Krasner Foundation / Artists Rights Society (ARS), New York
Jeopardy	Associated Press Carol Kaelson/ Jeopardy Productions Inc
Programming Cartoon	Kind permission Geekherocomic
Specification Cartoon	Credit Paragon Innovations

Chapter 12

Two Digital Brains Communicating	Photobank Gallery, www.shutterstock.com
Perpetual Motion from the 1600s	Robert Fludd's 1618 "Water Screw", Wikimedia, PD
Black Hole Malament-Holgarth Space	Crystal Graphics
Synapses and Tubulin	Crystal Graphics

Chapter 13

World Communication	Antartis, www.bigstock.com
IMAX	Louie Psihoyos/Corbis
Hologram	videodoctor, www.shutterstock.com

Chapter 14

Invention of Light Bulb, Thomas Edison	Corbis, Betmann
Steve Jobs Shows the iPhone	Corbis, Reuters
Stopwatch 60 seconds!	Studio 37, www.shutterstock.com
Paperclip Test	Illustrated by James Tagg
30 Things Test	Illustrated by James Tagg
Paperclip Test2	Illustrated by James Tagg
Eureka	KoS, Wikimedia, PD
Circle with Dot, Problem	Illustrated by James Tagg
Thinking Outside the Box	Illustrated by James Tagg
John Cleese, Video Link	Picture from www.shutterstock, links to World Innovation Forum, YouTube talk in iBook version and on website.
Sketch Test	Illustrated by James Tagg
Hard Drives	www.shutterstock.com
Harold Cohen and AARON	James Tagg at the Computer Museum
Harold Cohen and AARON	James Tagg at the Computer Museum
Old Polo	Generic Polo Photo
New Polo	Fingerhut, www.shutterstock.com

Chapter 15

Dilbert on Free Will	DILBERT © 1993 Scott Adams. Used By permission of UNIVERSAL UCLICK. All rights reserved.
Domino Toppling	(c) www.austriandominoart.com

Newton's Rings	Wikimedia
Wave Interference	Single image in Book. Slide show in iBook, Various; Wikimedia www.shutterstock.com, www.bigstock.com
Solvay Conference	Benjamin S. Couprie, Wikimedia, PD
Interferometer	Illustrated by James Tagg
Schrödinger's Cat	Dhatfield, Wikimedia, CC3
Polarized Glasses, Glare and No Glare	HUB, Wikimedia, CC3
Bell Test	Illustrated by James Tagg
Left and Right Socks	Hofmeester, Bigstock.com
Morse Signaling	Illustrated by James Tagg
Dawkins and Atheist Bus	Wikimedia, CC2
Fork in the Road	fivepointsix, www.bigstock.com
Left Hand Rule	Photograph by James Tagg
Orthogonal Sticks	Illustrated by James Tagg
M.C. Escher's "Waterfall"	© 2014 The M.C. Escher Company-The Netherlands. All rights reserved. www.mcescher.com
Kochen-Specker Cube	James Tagg modeled in Sketchup
Russian Dolls	Robyn Mackenzie, www.bigstock.com

Chapter 16

The HMS Beagle	Bettmann/Corbis
Nobel Prize Medal	Wikimedia, PD
Golden Hall, Sweden	vichie81, www.shutterstock.com
Pulitzer Prize Medal	Original Daniel Chester French, photo upload Katpatuka, Wikimedia, PD
First XPRIZE Award Ceremony	Kbh3rd, Wikimedia, CC3
Fields Medal	Stefan Zachow, Wikimedia, PD

Chapter 17

Omar Khayyam	Wikimedia, PD

Appendices

Leonardo da Vinci, Self Portrait	Wikimedia, PD
British Library	Diliff, Wikimedia, CC2.5
Experiment, ATLAS, CERN	xdrew, www.shutterstock.com
Panda	leungchopan, www.shutterstock.com
Conway and Kochen	Photograph courtesy of Princeton University's Office of Communications; Denise Applewhite, photographer
Looney Tunes "That's all Folks"	Wikimedia, PD

The Wikimedia Creative Commons Licenses 1, 2, 2.5 and 3 may be found at www.wikimedia.com. PD indicates a public domain.

In the case of items marked 'video' clicking on the image in the iBook or eBook will link to YouTube. The links are also available at www.jamestagg.com/videolinks for book readers.

Reading Room at the British Museum

"From the moment I picked your
book up until I laid it down I
was convulsed with laughter.
Some day I intend reading it."

Groucho Marx

Appendix 2

Bibliography

I am not resident at an academic institution, nor do I work for a large company with access to a broad range of journal subscriptions. I have a degree in Physics and Computer Science, so I am no layman. The modern web gives amateurs like me, easy access to enormous information resources that would only have been available from the finest University libraries even five years ago. Over time I have built up a personal library of books in the field, many of them ex-library copies which, by their date stamps, were never borrowed in their home universities!

I'm always skeptical of the enormous bibliographies found in the back of science books and whether they are ever read. If you want a pointer to the next books to read, here are some suggestions: *A Brief History of Time, The Man who Mistook his Wife for a Hat, The Emperor's New Mind, The Naked Jape, Gödel Escher Bach, Proust and the Squid, Logic, A Five Day Course in Thinking, Your Brain on Music, The 4% Universe, From Eternity to Here* and *Time.*

Journal	Cost
Scientific American	Some free articles, Membership $150 p.a.
Wikipedia	Free
The Economist	Free to search but a subscription needed for full articles, $200
Science: The American Institute for the Advancement of Science.	Annual Subscription $151
Mind	Annual Subscription, $200
Google Scholar	Free to search and often free to view, but some articles require membership of underlying services. From here you jump off into an endless series of journals too numerous to mention.
Google Books	Free, but you buy a lot of books!
Amazon	Some free material but again lot of book buying
JStor	Variable, based on area of interest
Arxiv.org	All the pre-prints of forthcoming papers. Invaluable. Free
SpringerLink	Article by article purchase, $35 each

Subscriptions and Sources

Chapter 1

Bellos, Alex. *Alex's Adventures in Numberland*. Bloomsbury Publishing PLC, 2010.

by Richard Roeper. *Urban Legends: The Truth behind All Those Deliciously Entertaining Myths That Are Absolutely, Positively, 100 Percent Not True*. Career Press, 1999.

Cairns-Smith, A. Graham. *Evolving the Mind: On the Nature of Matter and the Origin of Consciousness*. Cambridge University Press, 1996.

Dawkins, Richard. *The Magic of Reality: How We Know What's Really True*. Bantam Press, 2011.

Ericsson, K. Anders. "Attaining Excellence through Deliberate Practice: Insights from the Study of Expert Performance." *The Pursuit of Excellence through Education*, 2002, 21–55.

———. "Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains." *Academic Medicine* 79, no. 10 (2004): S70–81.

Ericsson, K. Anders, Ralf T. Krampe, and Clemens Tesch-Römer. "The Role of Deliberate Practice in the Acquisition of Expert Performance." *Psychological*

Review 100, no. 3 (1993): 363.

Fiske, John. *Introduction to Communication Studies*. 3rd ed. Routledge, 2010.

Franklin, Stan. *Artificial Minds*. MIT Press, 1997.

Gilovich, Thomas. *How We Know What Isn't So: Fallibility of Human Reason in Everyday Life*. Reprint. The Free Press, 1993.

Gregory, Robert J. *Psychological Testing: History, Principles, and Applications*. 6th ed. Pearson, 2010.

Hameroff, S. R. "Quantum Coherence in Microtubules: A Neural Basis for Emergent Consciousness?" *Journal of Consciousness Studies* 1, no. 1 (1994): 91–118.

Hameroff, Stuart R., and Alfred W. Kaszniak. *Toward a Science of Consciousness: The First Tucson Discussions and Debates*. MIT Press, 1996.

Harel, David. *Computers Ltd: What They REALLY Can't Do*. New Ed. OUP Oxford, 2003.

Hawkins, Jeff, and Sandra Blakeslee. *On Intelligence*. Reprint. Owl Books (NY), 2005.

Hofstadter, Douglas R. *Godel, Escher, Bach: An Eternal Golden Braid*. 20th Anniversary ed. Penguin, 2000.

Howell, E. *From Darness Light/Land of Stone/Shadow Worlds*. Centaur, 2010.

"IBM100 - Deep Blue." CTB14, March 7, 2012. http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/.

Ivancevic, Vladimir G., and Tijana T. Ivancevic. *Quantum Neural Computation*. Springer, 2010.

Jibu, Mari, and Kunio Yasue. *Quantum Brain Dynamics and Consciousness: An Introduction*. John Benjamins Publishing, 1995.

Jibu, M., S. Hagan, S. R. Hameroff, K. H. Pribram, and K. Yasue. "Quantum Optical Coherence in Cytoskeletal Microtubules: Implications for Brain Function." *Biosystems* 32, no. 3 (1994): 195–209.

Lahoz-Beltra, R., S. R. Hameroff, and J. E. Dayhoff. "Cytoskeletal Logic: A Model for Molecular Computation via Boolean Operations in Microtubules and Microtubule-Associated Proteins." *BioSystems* 29, no. 1 (1993): 1–23.

Malcolm Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Repr. Abacus, 2001.

Morris, Desmond. *Child: How Children Think, Learn and Grow in the Early Years*. Hamlyn, 2010.

———. *The Naked Ape: A Zoologist's Study of the Human Animal*. New edition. Vintage, 2005.

Neumann, John Von. *The Computer and the Brain*. 2nd Revised edition. Yale University Press, 2000.

Penrose, Roger. *The Large, the Small and the Human Mind*. New Ed. Cambridge University Press, 2000.

Přibram, Karl H., and Sir John Carew Eccles. *Rethinking Neural Networks: Quantum Fields and Biological Data*. Routledge, 1993.

"Return to Antikythera: Divers Revisit Wreck Where Ancient Computer Found." *The Guardian*, October 2, 2012. http://www.guardian.co.uk/science/blog/2012/oct/02/return-antikythera-wreck-ancient-computer.

Robinson, Ken, and Lou Aronica. *The Element: How Finding Your Passion Changes Everything*. Penguin, 2010.

Roebuck, Kevin. *Emotional Intelligence: High-Impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors*. Tebbo,

2011.
Sacks, Oliver. *An Anthropologist on Mars*. 4th ed. Picador, 2009.
———. *The Man Who Mistook His Wife for a Hat*. 1st ed. Picador, 1986.
Tucker, William H. *The Cattell Controversy: Race, Science, and Ideology*. University of Illinois Press, 2009.
Turing, Alan M. "Intelligent Machines." *Ince, DC (Ed.)* 5 (1992). http://isites.harvard.edu/fs/docs/icb.topic958294.files/lecture-00-handout.pdf.
Vitiello, Giuseppe. *My Double Unveiled: The Dissipative Quantum Model of Brain*. John Benjamins Publishing, 2001.
Whitehead, Alfred North, and Bertrand Russell. *Principia Mathematica - Volume One: 1*. Rough Draft Printing, 2009.
Winston, Robert. *The Human Mind and How to Make the Most of It*. New edition. Chartered Institute of Personnel and Development, 2006.
Wolf, Maryanne. *Proust and the Squid: The Story and Science of the Reading Brain*. Icon Books Ltd, 2008.
Wolfram, Stephen. *A New Kind of Science*. First Edition. Wolfram Media Inc, 2002.

Chapter 2

Carr, Jimmy, and Lucy Greeves. *The Naked Jape: Uncovering the Hidden World of Jokes*. Penguin, 2007.
Cobley, Paul. *The Communication Theory Reader*. 1st ed. Routledge, 1996.
Dawkins, Richard. *Unweaving the Rainbow: Science, Delusion and the Appetite for Wonder*. Reisssue. Penguin, 2006.
Hume, David. *An Enquiry Concerning Human Understanding*. New Ed. /. OUP Oxford, 2008.
Locke, John. *An Essay Concerning Human Understanding*. Abridged edition. Hackett Publishing Co, Inc, 1996.
Martin, Robert M. *There Are Two Errors in the the Title of This Book**. Rev. and Expanded Ed. Broadview Press Ltd, 2002.
Sacks, Oliver. *An Anthropologist on Mars*. 4th ed. Picador, 2009.
Tufte, Edward R. *The Cognitive Style of PowerPoint: Pitching Out Corrupts Within*. 2nd ed. Graphics Press, 2006.
Wiseman, Prof. Richard. *Quirkology: The Curious Science Of Everyday Lives*. 2nd ed. Pan, 2011.

Chapter 3

Borg, James. *Body Language: 7 Easy Lessons to Master the Silent Language*. 1st ed. Prentice Hall Life, 2008.
Brounstein, Marty. *Communicating Effectively for Dummies*. John Wiley & Sons, 2001.
by Seth Godin ; with a foreword by Malcolm Gladwell. *Unleashing the Ideavirus : How to Turn Your Ideas into Marketing Epidemics*. Free Press, 2002.
Darwin, Charles. *The Origin of Species*. New edition. Wordsworth Editions Ltd, 1998.
Fiske, John. *Introduction to Communication Studies*. 3rd ed. Routledge, 2010.
Morris, Desmond. *Peoplewatching: The Desmond Morris Guide to Body Language*. Vintage, 2002.
———. *The Human Zoo*. New edition. Vintage, 1994.
———. *The Naked Ape: A Zoologist's Study of the Human Animal*. New edition. Vintage, 2005.

Navarro, Joe. *What Every Body Is Saying: An Ex-FBI Agent's Guide to Speed-Reading People*. HarperCollins Publishers, 2008.

Ogilvy, David. *Ogilvy on Advertising*. New edition. Prion Books Ltd, 2007.

Roebuck, Kevin. *Emotional Intelligence: High-Impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors*. Tebbo, 2011.

Schirato, Tony, and Susan Yell. *Communication and Culture: An Introduction*. 2nd Revised edition. Sage Publications Ltd, 2000.

Taylor, Kathleen. *Brainwashing: The Science of Thought Control*. New Ed. OUP Oxford, 2006.

Winston, Professor Lord Robert. *Human Instinct*. New edition. Bantam, 2008.

Chapter 4

Derren Brown. *Tricks of the Mind*. Channel 4, 2006.

Gurney, Kevin. *An Introduction to Neural Networks*. CRC Press, 1997.

Hameroff, S. R. "Quantum Coherence in Microtubules: A Neural Basis for Emergent Consciousness?" *Journal of Consciousness Studies* 1, no. 1 (1994): 91–118.

Higbee, Kenneth L., and Ph.D. *Your Memory: How It Works and How to Improve It*. 2Rev Ed. Avalon Group, 2001.

Jibu, M., S. Hagan, S. R. Hameroff, K. H. Pribram, and K. Yasue. "Quantum Optical Coherence in Cytoskeletal Microtubules: Implications for Brain Function." *Biosystems* 32, no. 3 (1994): 195–209.

Jimmy Carr. *The Naked Jape : Uncovering the Hidden World of Jokes*. Michael Joseph, 2007.

Lahoz-Beltra, R., S. R. Hameroff, and J. E. Dayhoff. "Cytoskeletal Logic: A Model for Molecular Computation via Boolean Operations in Microtubules and Microtubule-Associated Proteins." *BioSystems* 29, no. 1 (1993): 1–23.

O'Brien, Dominic. *How to Develop a Brilliant Memory Week by Week: 52 Proven Ways to Enhance Your Memory Skills*. Duncan Baird Publishers, 2005.

Picton, P.D. *Neural Networks*. 2nd Revised edition. Palgrave Macmillan, 2000.

Siegelmann, Hava T. *Neural Networks and Analog Computation: Beyond the Turing Limit*. Birkhauser, 1998.

Winston, Robert. *The Human Mind and How to Make the Most of It*. New edition. Chartered Institute of Personnel and Development, 2006.

Wolf, Maryanne. *Proust and the Squid: The Story and Science of the Reading Brain*. Icon Books Ltd, 2008.

Chapter 5

Aaronson, Scott. *Quantum Computing since Democritus*. New York: Cambridge University Press, 2013.

Borges, Jorge Luis, and Andrew Hurley. *The Library of Babel*. Boston: David R. Godine Publisher Inc, 2000.

Chaitin, Gregory J. *Meta Maths: The Quest for Omega*. Atlantic Books, 2007.

Tolstoy, Leo. *War and Peace*. Ware: Wordsworth Editions, 2001.

Chapter 6

Ayer, A. J. *Language, Truth and Logic*. 2nd ed. Dover Publications Inc., 2002.

Boaler, Jo. *The Elephant in the Classroom: Helping Children Learn and Love Maths*. Souvenir Press Ltd, 2010.

Carroll, Lewis. *Lewis Carroll's Games and Puzzles*. 40th ed. Dover Publications Inc.,

1992.
———. *Symbolic Logic*. New issue of 1896 ed. Dover Publications Inc., 2000.
Crilly, Tony. *The Big Questions: Mathematics*. Quercus Publishing Plc, 2011.
Doxiadis, Apostolos, and Christos H. Papadimitriou. *Logicomix: An Epic Search for Truth*. First Edition. Bloomsbury Publishing PLC, 2009.
Goldrei, D.C. *Classic Set Theory: A Guided Introduction*. Chapman and Hall/CRC, 1996.
Hodges, Wilfrid. *Logic*. 2nd Revised edition. Penguin, 2001.
Martin, Robert M. *There Are Two Errors in the the Title of This Book**. Rev. and Expanded Ed. Broadview Press Ltd, 2002.
Newbery, John. *Logic Made Familiar and Easy: To Which Is Added a Compendious System of Metaphysics Or Ontology : Being the Fifth Volume of the Circle of the Sciences, &c. Published by the King's Authority*. BiblioBazaar, LLC, 2010.
Oechslin, Werner. *Byrne, Six Books of Euclid: Facsimile of the Famous First Edition of 1847*. Har/Pap. Taschen GmbH, 2010.
Russell, Bertrand. *Introduction to Mathematical Philosophy*. Reprint. Spokesman Books, 2007.

Chapter 7

Gleick, James. *Chaos: Making a New Science*. New edition. Vintage, 1997.
Griffeath, David, and Cristopher Moore. *New Constructions in Cellular Automata*. Oxford University Press, 2003.
"Mathematical Games - The Fantastic Combinations of John Conway's New Solitaire Game 'Life' - M. Gardner - 1970," June 3, 2009. http://web.archive.org/web/20090603015231/http://ddi.cs.uni-potsdam.de/HyFISCH/Produzieren/lis_projekt/proj_gamelife/ConwayScientificAmerican.htm.
Mitchell, Melanie. *Complexity: A Guided Tour*. OUP USA, 2009.
Weisstein, Eric W. "Elementary Cellular Automaton." Text. Accessed September 28, 2014. http://mathworld.wolfram.com/ElementaryCellularAutomaton.html.
Wolfram, Stephen. *A New Kind of Science*. First Edition. Wolfram Media Inc, 2002.

Chapter 8

Cantor, Georg. *Contributions to the Founding of the Theory of Transfinite Numbers*. Dover Publications Inc., 2003.
Clegg, Brian. *Brief History of Infinity: The Quest to Think the Unthinkable*. Robinson Publishing, 2003.
Cohen, Paul J. *Set Theory and the Continuum Hypothesis*. Dover Publications Inc., 2009.
Pica, Pierre, and Alain Lecomte. "Theoretical Implications of the Study of Numbers and Numerals in Mundurucu." *Philosophical Psychology* 21, no. 4 (August 1, 2008): 507–22. doi:10.1080/09515080802285461.
Pica, Pierre, Cathy Lemer, Véronique Izard, and Stanislas Dehaene. "Exact and Approximate Arithmetic in an Amazonian Indigene Group." *Science* 306, no. 5695 (October 15, 2004): 499–503. doi:10.1126/science.1102085.
Potter, Michael. *Set Theory and Its Philosophy: A Critical Introduction*. Clarendon Press, 2004.

Chapter 9

Chaitin, Gregory J. *Thinking About Gödel And Turing: Essays On Complexity 1970-2007*. World Scientific Publishing, 2007.

Franzén, Torkel. *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse*. A K Peters/CRC Press, 2005.

Gödel, Kurt. *On Formally Undecidable Propositions of "Principia Mathematica" and Related Systems*. New edition. Dover Publications Inc., 2003.

Goldrei, D.C. *Classic Set Theory: A Guided Introduction*. Chapman and Hall/CRC, 1996.

Hofstadter, Douglas R. *Gödel, Escher, Bach: An Eternal Golden Braid*. 20th Anniversary ed. Penguin, 2000.

Nagel, Ernest, and James R. Newman. *Gödel's Proof*. Rev. Ed. New York University Press, 2001.

Newton, Sir Isaac. *Principia*. Prometheus Books, 1995.

Penrose, Sir Roger. *Shadows Of The Mind: A Search for the Missing Science of Consciousness*. New edition. Vintage, 2005.

Potter, Michael. *Set Theory and Its Philosophy: A Critical Introduction*. Clarendon Press, 2004.

Russell, Bertrand. *Introduction to Mathematical Philosophy*. Reprint. Spokesman Books, 2007.

Sautoy, Marcus Du. *The Music of the Primes: Why an Unsolved Problem in Mathematics Matters*. New Ed. Harper Perennial, 2004.

———. *The Number Mysteries: A Mathematical Odyssey Through Everyday Life*. Fourth Estate, 2010.

Whitehead, Alfred North, and Bertrand Russell. *Principia Mathematica - Volume One: 1*. Rough Draft Printing, 2009.

Chapter 10

Copeland, B. Jack. *The Essential Turing*. Clarendon Press, 2004.

David, Hans T. *The New Bach Reader: Life of Johann Sebastian Bach in Letters and Documents*. New edition. W. W. Norton & Co., 1999.

Dennett, Daniel C., and Douglas R. Hofstadter. *The Mind's I: Fantasies and Reflections on Self and Soul*. Basic Books, 2000.

Dewdney. *New Turing Omnibus*. Reprint. Palgrave Macmillan, 2003.

edited by B. Jack Copeland. *The Essential Turing : Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life plus the Secrets of Enigma*. Reprinted. Clarendon Press, 2005.

Gallwey, W Timothy, and Barry Green. *Inner Game of Music*. 7th ed. Pan, 2003.

Hofstadter, Douglas R. *I Am a Strange Loop*. Reprint. Basic Books, 2008.

Levitin, Daniel J. *This Is Your Brain on Music: Understanding a Human Obsession*. Atlantic Books, 2008.

Penrose, Roger. *The Large, the Small and the Human Mind*. New Ed. Cambridge University Press, 2000.

Penrose, Sir Roger. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New Ed. Oxford Paperbacks, 1999.

Petzold, Charles. *The Annotated Turing: A Guided Tour Through Alan Turing's Historic Paper on Computability and the Turing Machine*. John Wiley & Sons, 2008.

Sacks, Oliver. *Musicophilia: Tales of Music and the Brain*. Reprint. Picador, 2008.

Singh, Simon. *The Code Book: The Secret History of Codes and Code-Breaking*. (Reissue). Fourth Estate, 2002.

Turing, Alan. "Checking a Large Routine." In *The Early British Computer Conferences*, 70–72. MIT Press, 1989. http://dl.acm.org/citation.

cfm?id=94952.
Turing, Alan M. "Can a Machine Think." *The World of Mathematics* 4 (1956): 2099–2123.
———. "Computability and Λ-Definability." *The Journal of Symbolic Logic* 2, no. 4 (1937): 153–63.
———. "Computing Machinery and Intelligence." *Mind*, 1950, 433–60.
———. "Computing Machinery and Intelligence." In *Computers & Thought*, 11–35. MIT Press, 1995. http://dl.acm.org/citation.cfm?id=216410.
———. "Intelligent Machines." *Ince, DC (Ed.)* 5 (1992). http://isites.harvard.edu/fs/docs/icb.topic958294.files/lecture-00-handout.pdf.
———. "Rounding-off Errors in Matrix Processes." *The Quarterly Journal of Mechanics and Applied Mathematics* 1, no. 1 (1948): 287–308.
Turing, Alan Mathison. "On Computable Numbers, with an Application to the Entscheidungsproblem." *J. of Math* 58 (1936): 345–63.
———. "Systems of Logic Based on Ordinals." *Proceedings of the London Mathematical Society* 2, no. 1 (1939): 161–228.
———. "The Chemical Basis of Morphogenesis." *Bulletin of Mathematical Biology* 52, no. 1 (1990): 153–97.

Chapter 11

Baxa, Christoph. "A Note on Diophantine Representations." *American Mathematical Monthly*, 1993, 138–43.
Blass, Andreas, and Yuri Gurevich. "Algorithms: A Quest for Absolute Definitions." *Bulletin of the EATCS* 81 (2003): 195–225.
Börger, Egon, Erich Grädel, and Yuri Gurevich. *The Classical Decision Problem*. Springer, 2001.
Carroll, Lewis. *Symbolic Logic*. New issue of 1896 ed. Dover Publications Inc., 2000.
Davis, Martin, Hilary Putnam, and Julia Robinson. "The Decision Problem for Exponential Diophantine Equations." *Annals of Mathematics*, 1961, 425–36.
Dyson, Verena H., James P. Jones, and John C. Shepherdson. "Some Diophantine Forms of Gödel's Theorem." *Archive for Mathematical Logic* 22, no. 1 (1980): 51–60.
Franzén, Torkel. *Godel's Theorem: An Incomplete Guide to Its Use and Abuse*. A K Peters/CRC Press, 2005.
Hodges, Wilfrid. *Logic*. 2nd Revised edition. Penguin, 2001.
Jr, Frederick P. Brooks. *The Mythical Man Month and Other Essays on Software Engineering*. 2nd ed. Addison Wesley, 1995.
Kurzweil, Ray. *How to Create a Mind: The Secret of Human Thought Revealed*. Penguin, 2012.
Matiyasevich, Yuri. *HILBERT'S TENTH PROBLEM: What Can We Do with Diophantine Equations?*. Accessed April 13, 2014. http://logic.pdmi.ras.ru/~yumat/Journal/H10history/H10histe.pdf.gz.
Minsky, Marvin Lee. *Computation*. Prentice-Hall Englewood Cliffs, 1967. http://cba.mit.edu/events/03.11.ASE/docs/Minsky.pdf.
Nagel, Ernest, and James R. Newman. *Godel's Proof*. Rev. Ed. New York University Press, 2001.
Penrose, Sir Roger. *Shadows Of The Mind: A Search for the Missing Science of Consciousness*. New edition. Vintage, 2005.
———. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New Ed. Oxford Paperbacks, 1999.

Reid, Constance. *Julia: A Life in Mathematics*. Washington, DC: The Mathematical Association of America, 1997.

Rice, Henry Gordon. "Classes of Recursively Enumerable Sets and Their Decision Problems." *Transactions of the American Mathematical Society* 74, no. 2 (1953): 358–66.

Ruohonen, Keijo. "Hilbertin Kymmenes Probleema." *Arkhimedes*, no. 1 (1972): 2.

Spolsky, J. "The Law of Leaky Abstractions," November 11, 2002. http://www.joelonsoftware.com.

Tegmark, Max. "The Mathematical Universe." *Foundations of Physics* 38, no. 2 (2008): 101–50.

Turing, Alan M. "Can a Machine Think." *The World of Mathematics* 4 (1956): 2099–2123.

Turing, Alan Mathison. "On Computable Numbers, with an Application to the Entscheidungsproblem." *J. of Math* 58 (1936): 345–63.

———. "Systems of Logic Based on Ordinals." *Proceedings of the London Mathematical Society* 2, no. 1 (1939): 161–228.

Wiles, Andrew. "Modular Elliptic Curves and Fermat's Last Theorem." *Annals of Mathematics-Second Series* 141, no. 3 (1995): 443–552.

Chapter 12

Burgin, Mark. *Super-Recursive Algorithms*. Springer, 2005.

Darwin, Charles. *On the Origin of Species: By Means of Natural Selection*. Dover Giant Thrift Ed. Dover Publications Inc., 2006.

Mitchell, Melanie. *An Introduction to Genetic Algorithms*. New edition. MIT Press, 1998.

Siegelmann, Hava T. *Neural Networks and Analog Computation: Beyond the Turing Limit*. Birkhauser, 1998.

Syropoulos, Apostolos. *Hypercomputation: Computing Beyond the Church-Turing Barrier*. Softcover reprint of hardcover 1st ed. 2008. Springer, 2010.

Chapter 13

Shannon, C.E., and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

Chapter 14

Boden, Margaret A. *The Creative Mind: Myths and Mechanisms*. 2nd ed. Routledge, 2003.

Bono, Edward de. *How to Have Creative Ideas: 62 Exercises to Develop the Mind*. Vermilion, 2007.

———. *Lateral Thinking: A Textbook of Creativity*. Penguin, 2009.

———. *Six Thinking Hats*. Penguin, 2009.

Coyle, Daniel. *The Talent Code: Unlocking the Secret of Skill in Maths, Art, Music, Sport, and Just About Everything Else*, Random House Books, 2009.

Davis, Ronald D., and Eldon M. Braun. *The Gift of Dyslexia: Why Some of the Brighest People Can't Read and How They Can Learn*. 3rd Revised edition. Souvenir Press Ltd, 2010.

Edward De Bono. *How to Have Creative Ideas : 62 Exercises to Develop the Mind*. 1 Aufl. Vermilion, 2007.

Isaacson, Walter. *Steve Jobs: The Exclusive Biography*. Little, Brown, 2011.

McCandless, David. *Information Is Beautiful*. Collins, 2010.

Robinson, Ken. *Out of Our Minds: Learning to Be Creative.* 2nd Edition. Capstone, 2011.

Robinson, Ken, and Lou Aronica. *The Element: How Finding Your Passion Changes Everything.* Penguin, 2010.

Winston, Professor Lord Robert. *Bad Ideas?: An Arresting History of Our Inventions.* Bantam, 2011.

Chapter 15

Bell, John S., and others. "On the Einstein-Podolsky-Rosen Paradox." *Physics* 1, no. 3 (1964): 195–200.

Conway, John H., and Simon Kochen. "The Strong Free Will Theorem." *Notices of the AMS* 56, no. 2 (2009): 226–32.

Conway, John, and Simon Kochen. "Reply to Comments of Bassi, Ghirardi, and Tumulka on the Free Will Theorem." *Foundations of Physics* 37, no. 11 (2007): 1643–47.

Dennett, Daniel Clement. *Brainstorms: Philosophical Essays on Mind and Psychology.* 8. MIT Press, 1981.

Dennett, Danile C. *Kinds of Minds: Toward an Understanding of Consciousness.* Basic Books, 2008.

Ekert, Artur K. "Quantum Cryptography Based on Bell's Theorem." *Physical Review Letters* 67, no. 6 (1991): 661.

"EPR Paradox." *Wikipedia, the Free Encyclopedia*, September 15, 2014. http://en.wikipedia.org/w/index.php?title=EPR_paradox&oldid=625734938.

Gilovich, Thomas. *How We Know What Isn't So: Fallibility of Human Reason in Everyday Life.* Reprint. The Free Press, 1993.

Gisin, Nicolas. "The Free Will Theorem, Stochastic Quantum Dynamics and True Becoming in Relativistic Quantum Physics." *arXiv Preprint arXiv:1002.1392*, 2010. http://arxiv.org/abs/1002.1392.

Goldstein, Sheldon, Daniel V. Tausk, Roderich Tumulka, and Nino Zanghì. "What Does the Free Will Theorem Actually Prove." *Notices of the AMS* 57, no. 11 (2010): 1451–53.

Hawking, Stephen, and Leonard Mlodinow. *The Grand Design: New Answers to the Ultimate Questions of Life.* Bantam Press, 2010.

Heywood, Peter, and Michael LG Redhead. "Nonlocality and the Kochen-Specker Paradox." *Foundations of Physics* 13, no. 5 (1983): 481–99.

Huang, Yun-Feng, Chuan-Feng Li, Yong-Sheng Zhang, Jian-Wei Pan, and Guang-Can Guo. "Experimental Test of the Kochen-Specker Theorem with Single Photons." *Physical Review Letters* 90, no. 25 (2003): 250401.

Russell, Bertrand. *The Problems of Philosophy.* 2nd ed. Oxford Paperbacks, 2001.

Tumulka, Roderich. "Comment on 'the Free Will Theorem.'" *Foundations of Physics* 37, no. 2 (2007): 186–97.

Zhang, Yong-Sheng, Chuan-Feng Li, and Guang-Can Guo. "Quantum Key Distribution via Quantum Encryption." *Physical Review A* 64, no. 2 (2001): 024302.

Chapter 16

Land, George, and Beth Jarman. *Breakpoint and Beyond: Mastering the Future - Today.* Reprint. HarperBusiness, 1993.

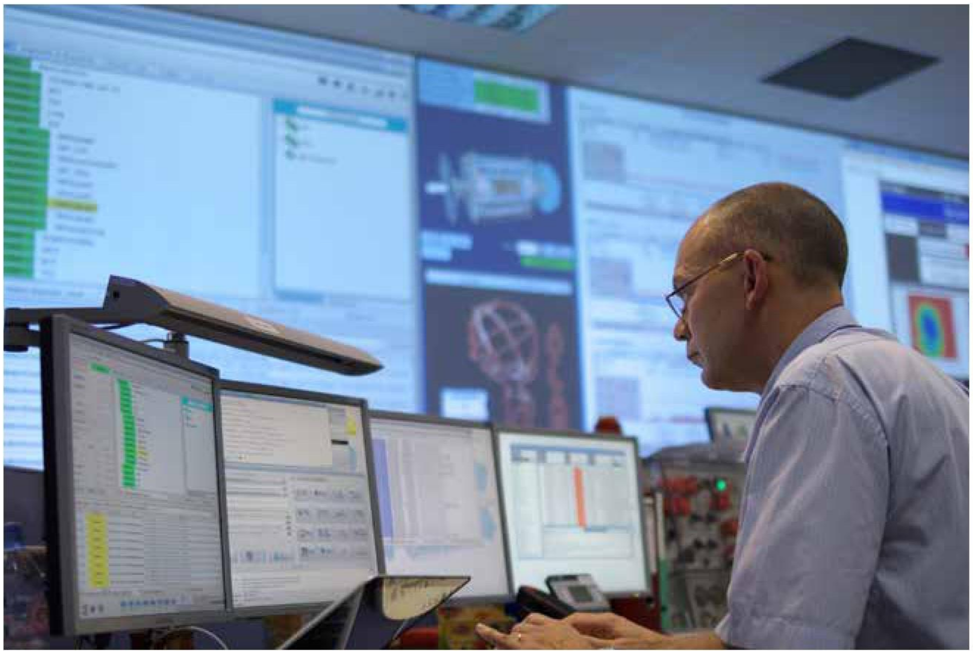
Lloyd, John, and John Mitchinson. *QI: The Second Book of General Ignorance.* Faber and Faber, 2010.

Michael Brooks. *13 Things That Don't Make Sense : The Most Intriguing Scientific Mysteries of Our Time*. Profile, 2010.

Reason Special Interview with Roger Penrose, 2008. http://www.youtube.com/watch?v=xiYDc1LA0I4&feature=youtube_gdata_player.

Reason Special Interview with Roger Penrose, 2008. http://www.youtube.com/watch?v=xiYDc1LA0I4&feature=youtube_gdata_player.

Tegmark, Max. "The Importance of Quantum Decoherence in Brain Processes." *arXiv:quant-ph/9907009*, July 5, 1999. doi:10.1103/PhysRevE.61.4194.



ATLAS, CERN

"Good one, publish."

Milliken

Appendix 3

Puzzles and Experiments

In this book I have suggested some experiments for you to undertake, and posed some puzzles to solve. You can participate in the experiments or see the answers by going online and checking my website at www.jamestagg.com

Jeopardy Answers from Chapter 1.

Answer 1. Watson answered, Gestate.

Answer 2. Watson answered. Who is Bram Stoker.

Answer 3. The answer is Chicago but Watson answered "Toronto?????", the question marks indicating it was very doubtful of the answer.

Panda, Eats Shoots

“Parenthetical remarks (however relevant) are unnecessary.”

Frank L. Visco,
How to Write Good

“Defining what we mean by a robot is hard to do. I know now when I see one but that definition works for anything, even pizza.”

Mike Gregory

Appendix 4

Conventions in the Book

The use of *italics* is intended to give accent or indicate the title of the published work, such as a book or movie. Inverted Commas should read, 'so called'.

The use of the 'Oxford Comma' – a comma after 'and' – gives a longer break in a sentence to aid better understanding. You will find both 'and' and, 'and,' used in this book deliberately.

The use of 'their' substitutes for he or she, and is a convention I believe will supersede, 'he' or 'she' and, 'him' or 'her'.

The puzzles in this book are all available to download, so you don't need to deface the book. Feel free to deface it if you wish. Buy the electronic version and save trees.

Some of this book is historical or factual, while much is highly speculative. I have tried to indicate where ideas are controversial and where they are matters of accepted science. However, my experience is that much of what you are taught tends to be a gross generalization or even plain wrong. If you treat facts with a degree of skepticism, you will find this keeps you in good stead.

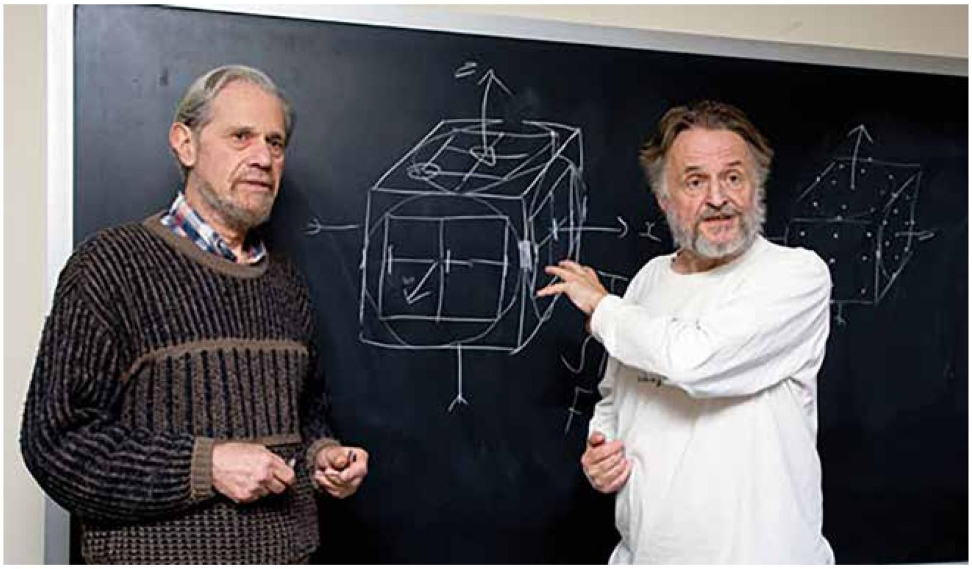
As someone who is highly dyslexic, there will be errors. Please email me the ones you find, and recommend the book to your friends so that I can afford to print a second edition with your corrections.

I have avoided the use of equations and mathematics, so you can see the flow of the philosophical argument. I also avoid a strongly historical narrative, but if you enjoy the history of science I recommend, *The Missing 4%*, and *Quantum* for a really clear exposé of the issues around quantum mechanics, relativity and cosmology, Robert Winston's *Bad Ideas* for a more thorough history of language and writing, and *The Trouble with Physics* for the recent history of particle Physics. Regarding the brain, anything by Oliver Sachs is a winner. *Your Brain on Music* is a good primer on the theory of artistic thought and music, and *Proust and the Squid* is a good discussion of dyslexia and learning.

The digits for the sequence of the Smallpox virus in Chapter Eight come from the varicella-zoster virus, a similar pathogen. The first few letters could be the same, but I hope it is impossible for anyone to disprove this as you would have to break into the CDC to do so.

John Masters is not the real name of the US officer who spoke in Kabul. That name was never released to the press.

This book was originally written using Microsoft Word on a MacBook Pro. It was then typeset for print using a variety of ePublication tools, including iBooksAuthor and Adobe InDesign. You can find a website for the book at www.jamestagg.com. It is published using WordPress. All the links in the book should be maintained there just in case there is link atrophy for the printed version of the book. Feel free to comment using the blog, Twitter, Facebook or email me.



Conway and Kochen

"It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong."

Richard P. Feynman

Appendix 5

Index of Theorems

I have explained a number famous theorems in this book along with some ideas of my own. Here is a list of the most notable.

1. **Mind over Computer**
   Turing Test – How to tell whether a computer is intelligent, even though we cannot agree on a definition of intelligence.
   Flynn Effect – The observation that human intelligence appears to be improving over time.
   Kurzweil Singularity Proposal – The conjecture that Moore's Law will result in computers acquiring near infinite power relatively soon.
   Lucas Argument – Minds are in a different class to computers as they are not limited by the incompleteness theorem.
   Humor Hypothesis† – Humor and jokes are a display of non-computable intelligence.

2. **Understanding**
   Tufte's Assertion – That communication of understanding exceeds the capability of many presentation tools, particularly PowerPoint.
   Chinese Room – John Searle's paradox challenging the idea that understanding can be mechanically simulated.

3. **Body Language & Banter**
   Communication Hypothesis† – That face to face communication is more powerful in some real physical sense than symbolic communication.
   7-38-55 Rule – Mehrabian's observation that the emotional content of communication is 7% words, 38% tone of voice and 55% body language.

4. **The Brain**
   Penrose-Hameroff Conjecture – Brains are quantum-gravity computers using tubulin as the mediator.

5. **Knowledge**
   Information Continuum Hypothesis† – Information is finite, but understanding and knowledge are different in nature.
   Infinite Monkeys Hypothesis – The paradox that says monkeys could type *Hamlet* given enough time and paper.
   Cat Experiment† – Finding my house cat can use our computer, and may be more creative than monkeys!

6. **Kittens & Gorillas**
   Feynman Proof – A proof that uses the lack of an evolved species within an evolutionary niche to disprove the existence of something; in this case polywater.
   The Infinity of Primes – Pythagoras' proof there are an infinity of prime numbers without needing the concept of a number.

7. **Complexity & Chaos**
   The Butterfly Effect – The proposal by Edward Lorenz that tiny effects can multiply up into enormous results.
   P≠NP – That non-deterministic polynomial problems can never be solved in polynomial time and are, therefore, beyond the capability of any imagined computer.
   The Hawking-Bekenstein Turk† – Although there is a perfect chess machine, it would collapse space-time to a black hole were it to exist.

8. **∞**
   Continuum Hypothesis – Does anything come between the first two infinities; counting numbers and the continuum of real numbers.

Cantor's First Infinity Theorem – The infinite plane is the same infinity as the infinite line.

9. **Known Unknowns**

Gödel's Incompleteness Theorem – Mathematics involving simple logic is incomplete.

Gödel's Completeness Theorem – First order logic is complete.

Hilbert's Completeness Theorem – Geometry is complete.

10. **Turing's Machines**

Entscheidungsproblem – The decision problem has no solution.

Turing Thesis – All computers, once sufficiently powerful, are equally powerful.

Non-Computability of Music† – That general musical compositions are non-computable.

Non-Computability of Creativity† – That general artistic creativity is non-computable.

11. **Software**

Brooks' Law – Adding resource to a late project makes it later.

Law of Leaky Abstractions – However good the attempt to abstract complexity, complexity has a habit of leaking through.

Software is Created† – writing software is a non-computable, creative task.

Bug Hypothesis† – Bugs are an inevitable consequence of trying to generalize software by mechanical means.

12. **Hyper-Computing**

Adaptive Recurrent Neural Network Hypothesis – Hava Siegelmann's proposal that ARNNs are capable of super-Turing computation.

13. **Hyper-Communication**

Bandwidth Conjecture† – Person-to-person communication has infinite bandwidth and is non-symbolic.

14. **Creativity**

Creativity Hypothesis† – That all creative endeavor is a non-computable skill, analogue to theorem discovery.

Wallas Model – A conceptual model for the way humans think creatively.

Creative Survival Advantage† – That creativity has evolved to give us a survival advantage at all stages of evolution, not just when we evolve to the point where mathematical intuition is significant.

15. **Free Will**

Laplace Daemon – A conceptual supreme being able to infer the future and the past from one snapshot of space-time and the laws of nature.

EPR Paradox – Twin particles would transmit information faster than the speed of light if quantum mechanics was to be believed.

Schrödinger's Cat – The paradox that a cat might be both alive and dead at the same time until observed.

Bell Inequality – Quantum mechanics forbids local hidden variables in a testable way. The test succeeds.

Kochen-Specker Paradox – Particles cannot know their settings before measurement.

Free Will Theorem – Nothing in the past light cone of the Universe causes a particle to choose its spin upon measurement.

Non-decryptable Universe† – The laws of physics mean the Universe is non-computable in principle.

Free Will Universe† – The laws of mathematics mean the Universe is non-computable and therefore has Free Will.

Russian Doll Conjecture† – Since humans are creative and non-deterministic and, in a sense, they run upon the hardware of the Universe, the hardware of our Universe must also be non-deterministic.

16. **The Quest for Knowledge**

Technology Hypothesis† – The extended strong anthropic principle that the Universe must have creative beings within it to uniquely define it.

17. **The Future**

Creative Non-Singularity† – The future is non-computable and therefore any increase in computer power, however great, will never achieve a creativity singularity.

† The symbol marks items proposed for the first time within this book. If I have missed a previous publication, please feel free to write to me and I will amend a future version.

Index

Symbols

∀∃∀ 249
∃ 249
3D chip 21, 22
3D printing 20, 222
7%-38%-55% rule 84
50 First Dates 12

A

AARON 310
Aaronson, Scott 168
abstraction 267
Academy Awards 365
Accidental Complexity 231
actin 98
Activision 290
Adams, Douglas 17, 128, 225, 238, 305, 314
Adams, Scott 296
Adaptive Recurrent Neural Network 280
Adleman, Leonard 165
Advanced Course in Design, Manufacturing and Management x
Afghanistan 54, 90
A Five-day Course in Thinking 297
age and memory 120
AIDS 158
Aleph 1 175
Alexander The Great 149
algebra 248
algorithm 212, 238, 239, 243, 247, 249, 251, 257
  history of 6
alibi 153
Al-Khwarizmi 6
Allman, Eric 265
Alternative Uses Task 299
Alvarez, Luis 30
Amazoncom 308
Amazon rainforest 27, 181
Amedi, Amir 13
Amherst 280
amygdala 11, 31
analysis 204
Analytical Engine 15, 16, 223–224
A New Kind of Science 173
anthropic principle 322
Anti-Ballistic Missile Treaty 79
Antikythera 15
Apple 56, 68, 224, 306
Arab world democracy 82
Arafat, Yasser 81, 82
Archimedes 343
Aristotle 149
ARM 224
ARNN 280
art
  being appreciated 142
  creativity and 304
Artamène 129
Art of Fugue 259
ASCII 203
Asimov, Isaac 4
Association of Computer Machinery 367

astrological clock 39, 42
asynchronous logic 22
Atkins v Virginia 28
ATM 233
ATP 98
audio field 291
audio processing system 114
auditory cortex 31
Australian English 86
Australian Outback 28
avatar xii
axiom 198
  Peano axioms 199, 205

B

Babbage, Charles 16, 223
Bach, JS 7, 259
background context 86
Bader, Douglas ix
bandwidth 288, 292
Barber paradox 155
Barrie, JM 128
Barrymore, Drew 12
Battleship 144
Baxa, Christoph 251
Beardsley, Dick 192
Beatles 102
Beijing 88
Bekenstein, Jacob 6, 132
Bell, Alexander Graham 18
Bell Corporation 216
Bell, John 322, 330, 333
Bell test experiment 333, 343
Berkeley University 248
bifocal glasses 152
Big Bang xii, 339
Big O 164
binary logic 150
Binsted, Kim 310
BIOS 225
Blackadder 266
Black Box experiments 67
black hole 132, 279
Bletchley Park 215
blind sight 12
Block Universe Hypothesis 317
body language 84, 288
Bohr, Niels 372
Bolt, Usain 31
bone cancer 20
Boolean logic 150, 152
Boole, George 150
Bootstrap 225
Borders 308
Borland 237
bosons 343
Bowie, David 151
brain 1, 11
  accidents 11
  aging 120
  amygdala 11
  anatomy 108
  and plasticity 13
  as a computer 13, 14, 99, 120
  as an exchange 18
  auditory cortex 31
  color perception 110
  digesting starch 118
  electrical pulse 98
  emotions 116
  glucose use 117
  hearing 113
  hippocampus 11, 32
  imaging 99, 102
    3D virtual 103
    MRI 104
    PET 107
    seeing thought 108
  learning 35–38
  memory 11, 14
  meninges 97
  motor cortex 31
  non-computable processes 373
  noninvasive imaging techniques 11
  organized like a filing cabinet 12
  quantum effects 50, 283
  quick tour 108
  scanners 46
  stroke damage 12
  super-Turing 294
  thinking 117
  visual agnosia 12
brain damage 97
brainstorming 298
Branson, Richard 24
Bricklin, Dan 237
bridge 134
British General Strike 211
Britten, Benjamin 305
Brooks, Fred 229, 230, 231, 237

Brooks' Law 230
bubble sort ballet 165
Buschkuehl, Martin 33
Bush, George W 194
butterfly
  creating tornados 173
Byrd, William 259
bytes 129

C

calculating machines 15
California Institute of Technology (Caltech) xii
Call of Duty 290
Cambridge University 17, 72, 195, 221
  King's College 211
  Mathematical Bridge x
  Trinity College 193
  Wolfson College ix
CaMKII 119
Candidate for a Pullet Surprise 139
Cantonese 87
Cantor, Georg 179
Cantor's theorem 221
Carey, Maria 113
Carroll, John 29
Carroll, Lewis 135, 140, 149, 148–150, 248, 356
Casio synthesizer 311
CAT scans 13, 102
Cats Creation 145
Cattell, Raymond 29
CERN 333, 335
Chaitin, George 188
Chalmers, David 39
Champollion, Jean-François 89
chaos 171
CHC theory 29
checklist 152
Cheshire Cat 149
chess 32, 163
  perfect chess-playing machine 6
chimpanzee and typewriter 127
China 15
Chinese 86, 129
Chinese Room, Searle's 65, 64–67, 222
Christensen, Clayton 306
Christie, Agatha 260
chunking 91
Church, Alonzo 212
Churchill, Winston 17, 213
Chutzpah 86
ciphers 215, 218
circle free 243
Clauser, John 333
Clay Mathematics Institute 167, 196, 225
Cleese, John 270, 303
clocks 42
  astrological clock 39, 42
  modern computers and 43
code breakers 211
codes 214
  and children's games 214
  and code books 215
  and code breakers 214, 305
  and Enigma machine 215, 305
  and one-time pad 216, 217, 218
Cohen, Harold 308, 311
COIN dynamics 53, 55, 90
Cold War 81, 213
color perception 110
comedy 92, 94
  as survival skill 94
communication
  audio field 291
  background context 86
  bandwidth 288, 292
  body language 288
  digitization 289, 293
  earliest recorded 87
  emails 81
  face to face xii, 81, 83, 293
  hyper-communication 285–294
  nonverbal 84
  of objects 90
  scripts and symbols 87
  symbolic 87, 293
  telephone 81
Compaq 307
compatibilism 41, 315
compiler 231
complexity
  accidental 231
  essential 231
complexity hierarchy 168
compositions 8
computer xi
  as human 11, 26, 46
  brains 14, 120
  bugs 262

  chess-playing 5
  communicating 90
  consciousness 14, 45
  crash 225
  creativity and 132, 257, 309
  Deep Thought 17
  first programmable computing machine 16
  generating random numbers 189
  historical convention 227
  infinite computing power 261
  Japanese characters 88
  limitless computing power 277
  logic 150
  logic gates 18, 22, 121
  logic limit 250
  military 80
  music and 258
  non-computable solution 75
  origins 15
  pattern matching 49
  personal 307
  programmed to learn 38
  quantum computers 277
  random numbers and 44
  sense of humor 25
  silicon chip 20
  symbolic communication 293
  synchronous logic 22
  understanding of 71
computer game 32
computing
  exponential growth 20
concentration 115
Confucius 162
Connelly, Jennifer 151
consciousness 45, 124
Conseil Européen pour la Recherche Nucléaire 333
Contact 68
convergent thinking 300
Conway, John 173, 343
Conway's Game of Life 173
Cope, David 7
Copenhagen interpretation 327
Copernicus 42
Cormack, Allan 102
counter factual experiments 282
counting system 181
Coyle, Daniel 37
Craddock, Travis 50, 119, 124, 283
Crash program 227
creative thinking 1
creativity 30, 49, 295–312, 297
  art and 304
  being appreciated 142
  computer and 132, 257, 309
  convergent thinking 300
  design tradeoff 311
  divergent thinking 298
  Eureka moment 302
  incubation 301
  innovator's dilemma 306
  intimation 301
  John Cleese on 303
  knowledge and 141
  mathematical creativity 309
  mechanical steps 141
  non-linearity of 311
  preparation 301
  process 133, 270, 305
  process versus 312
  reward for 308
  science of 301
  sparking creativity 305
Crete 89
cryptography 215
  quantum 217
CSI 102
Cuneiform 87
Curtis, Richard 266
cypher 214

D

Dahl, Roald 140
Damadian, Raymond 104
Daniel-Constantin Mierla 265
Danziger, Daniel 335
Dark Ages 16
Darwin, Charles 197, 305, 356
da Vinci, Leonardo 374
Davis, Martin 243, 248, 258
Dawkins, Richard 338–339
D-Day 213
dead code elimination 250
de Bono, Edward 297
Decision Problem 196, 212, 219
decohering 282
Dedekind 182
Deep Blue 5

Deep Thought 17
de Fermat, Pierre 75
Dell 307
democracy
  Arab world 82
De Morgan, Augustus 350
Demotic 89
dendrites 98
Dennett, Daniel xi, 46, 133, 260, 351
Der Spiegel 34
Descartes 70
design tradeoff 311
determined universe 207
determinism 41–46, xi, 315, 316, 351
  free will and 315
Deutsch, David xi, 56, 293, 309
Dexter, Colin 260
Dick, Philip K 324
diffusion MRI 106
digital art 129
Digital Equipment 307
digitization 289
  of life 293
Dijkstra, Edgar 3
Dilbert 313
Diophantine equations 238, 239, 240, 249, 251
Diophantus 239
divergent thinking 298
DNA 104, 123
domino toppling 316
Doyle, Arthur Conan 148
Doyon, Julien 117
drag and drop 237
DVD 289
D-Wave 21
dyslexia 90, 139

E

Edinburgh Festival 288
Edison, Thomas 295, 296, 298
EEG 32
Egyptian 86, 89
Einstein, Albert 30, 34, 47, 64, 117, 178, 180, 210, 212, 298, 319, 327, 332, 333, 347, 348
  brain 97
Eliza 25
Elton, Ben 266
EMI 102
Emil Post's Word Problem 258
emotions 116
encryption 165, 218, 277
  RSA encryption 165
Encyclopedia Britannica 8
English 86, 89, 129
ENIAC 223
Enigma 211, 212, 215, 218
Entscheidungsproblem 211, 219, 221
EPR paradox 332
Equitable Center 5
equivalence 200
Ericsson, Anders 37
Ernő 169
Escher, MC 346
  and Penrose Steps 51
Essential Complexity 231
Euclid's proof 156
Euler 157, 202
Eureka moment 302
Everett, Hugh 328
experiments 1
exponent 241
expression analysis 84
eyes 108
  color perception 110
  fovea centralis 112
  resolution of 112

F

Facebook 71, 271
face-to-face interaction 83
false paradox 155
Fermat's Last Theorem xi, 225, 241, 242, 251, 253, 254, 259, 261, 275, 278, 351, 368
Fermat's puzzle 225
Fermilab 293
Feynman, Richard 58, 70, 157, 239, 306, 326, 364, 399
Feynman's proof 157
Fields Fellowship 368
Fields Medal 368
Finland 54
FitzGerald, Edward 372
Florida State University 37
flowchart 244
Fluid Concepts & Creative Analogies 49
fluid intelligence 29, 33
fluorescent dyes 100

Flynn Effect 33
Flynn, James 33
Formalism 193
Four Color Conjecture 251
fovea centralis 112
f-PET 107
fractals 329
Franklin, Benjamin 152
Frankston, Bob 237
Freedman, Stuart 333
FreeSWITCH 265
free will 41, 313–354
  determinism and 315
  God and 339
  particles 349
  Schrödinger's cat 325
  simple theorem 337
  The Free Will Theorem 343
  twin particle experiment 331
  uncertainty 318
Frege, Gottlob 155
French 90
French Academy of Science 238
Fritz, chess program 5
Frost, Robert 96
future 373
futures market 55

G

gadolinium 106
Garden of Eden 339
gate parity point 18
Gates, Bill 236
gears 42
Geiger counter 326
geometry 179
German 86, 129
Gettysburg Address 57
'G' factors 29
ginormous 131
Giseng, Nicolas 335
Gladwell, Malcolm 37
Glass, Philip 8
Gleick, James 171
glucose 117
God
  free will and 339
Gödel Escher Bach 49
Gödel, Kurt xi, 141, 193, 201, 221, 246
Gödel limit 207
Gödel numbers 203
Goldbach's Conjecture 157
Golden Pineapples 365
Good Will Hunting 257
Google 20, 253, 271, 308, 367
Gorbachev, Mikael 79
Göttingen University 193
grade inflation 32
Graham, Martha 78
Grand Masters 5
Grand Theft Auto 91
Grantchester 221
Greece 15, 156
  and ancient Greeks 18
Greek 87, 89, 149, 297
Greek tragedy 87
Gregory, Mike 396
Grieg 259
Group Intelligence 30
Grove, Andy 18
Grover's algorithm 277
guess xii
Guilford, JP 299
Gulf Stream 152
Gurdon, Sir John 30

H

Haltcrash 244
Halting Flowchart 245
Halting Problem 243, 251, 258, 259
Halting Program 225, 244
Halting Question 243
Hameroff, Stuart 119, 122, 282
Hamlet 129, 254
Hammond, Richard 97
Hampton Court Palace 42
Hard disk drives 306
Harrison, John 365
Harry Potter 92, 306
Harvard Business School 306
Harvard University 224, 307
Hawking Bekenstein bound 6
Hawking, Stephen x, xvi, 6, 132, 306, 326, 339
hearing 113
heat-sight 102
Hebrew 86, 87
Hebrew University of Jerusalem 13
Heisenberg's uncertainty principle 318, 335

Henry VIII 42
hieroglyphics 89
Higgs Boson 357
Hilbert, David 182, 193, 194, 238
Hilbert Problems 196, 221
Hilbert's 10th Problem 243, 253
Hilbert's Hotel 182
Himalayas 55
hippocampus 11, 32
Hitler, Adolf 213
Hoane, Joe 5
Hodges, Wilfrid 150
Höfði House 79, 80
Hofstadter, Douglas xi, 49, 310
Hogarth, Mark 280
Hole in the Wall Project 35
Holmes, Sherlock 148
hologram 290
Hong Kong 88
horizontal abstraction 266
Horn, John 29
Hounsfield, Sir Godfrey 102
Howell, Emily 7, 310
Hubble 348
Hulme, David 41
hunters and spears 181
hyper-communication 285–294, 293
  audio field 291
  bandwidth 288, 292
  digitization 289, 293
  reality 289
  symbolic communication 293
hyper-computing 273
hypercube 241–244
hypotenuse 242

I

IBM 207, 224, 306
  and Watson 8
  Watson Research Laboratory 5
Ig Nobel Prize 365
imaging 99
IMAX theatre 287, 290, 293
Inception 50
inconsistency
  in mathematics 204
inconsistency defense 207
incubation 1, 301
Indiana University 49
indirect proof 153, 158, 244
infinity , 179, 280
  history of 179
  how to count 180
  larger than infinity 185
Infinity Hotel 182
infrared light 102
innovator's dilemma 306
insight 2
inspiration 2
instinctive reactions 109
Institute of Advanced Mathematics 212, 258
Intel 18, 224, 367
intelligence 25
  fluid 29, 33
  'G' factors 29
  grade inflation 32
  human vs computing 75
  physical basis of 30
  quantitative numerical skills 30
  static 32
  time 29
  vision 29
interaction 85
  face-to-face 83
interferometer 326
International Congress of Mathematicians 196
International Mathematical Union 368
internet
  encryption 165, 277
internet protocol 267
intimation 2, 301
intuitive thinking 2
IP 267
iPad 319
iPhone 22, 238, 297
iPod 289
IQ 27, 33, 120
  IQ Test 27
Iraq 87
Iraq war 80
Irvine 280
ISABEL 14
iTunes 130

J

Jabberwocky 135, 136
Jaeggi, Susanne 33
Japanese 87

Jape 310
Jefferson, Thomas 308
Jeopardy 9
Jessie 145
Jobs, Steve 47, 68, 296–297, 297, 306
joke 93
  world's funniest 94
Joke Analysis and Production Engine 310
Jones, JP 251

K

Kahn, Philip 237
Kamailio 265
Kasparov, Garry 5, 34
  and Deep Blue 4–6
Kelvin, Lord 180
Kerr Metric 280
Khayyám, Omar 371
King's College 17, 211
Kish, Daniel 13
knitting 117
knowledge 8
  analysis 204
  creating 141
  creativity and 141
  difficulty discovering 204
  discovery of 142
  nature of 193
  search for 140
Kochen, Simon 343
Kochen Specker 344
Kochen-Specker Cube 347
Kochen-Specker paradox 345
Königsberg Bridges 202
Königsberg University 202
Kronecker 178
Kurzweil, Ray 20, 261
  and Moore's Law 19

L

lambda calculus 258
language 129
  body 84
Laplace, Pierre-Simon 317
lateral thinking 297
lava lamp 45
Lavarand 45
Law and Order 87
learning 35–38
Leibniz 319
Leicestershire 89
liar's paradox 154
library
  knowledge 143
light 100
  spectrum 102
lightning rod 152
Linear-a 89
Linear-b 89
Linux 265
Liszt, Franz 7, 129
Loch Ness Monster 142
Loch Ness Monster Song 138
Loebner prize 72
logic 149
  binary 150
  Boolean 150, 152
  checklists 152
  for computers 150
  limit 247
  purpose of 156
  reduction to the absurd 152
  Stoic 150
Logic 150
logic gates 18, 22, 121
logic limit 262
London Bridge 42
London marathon 192, 202
London Mathematical Society 15, 238
London School of Economics 301
London Science Museum 16
long multiplication 240
Lorenz Attractor 173
Lorenz, Edward 170, 172
Lotus Corporation 237
Lucas argument 205
Lucas, JR xi, 205
Lucas-Penrose argument 205

M

Madam Tussaud 75
magnetic fields 105
Magnetic Resonance Imaging 105
Makanin, Gennadií 258
Malament, David 280
Manchester University 236
Mandarin 87
Mandelbrot diagram 174
Mandelbrot Set 161
Manhattan Project 239

Marx, Groucho 153, 382
Massachusetts 89
Massachusetts Institute of Technology (MIT) 25, 165, 168, 307
Masters, John 54
mathematical proofs 156
mathematical theorems xi
mathematicians
  Polish 211
mathematics 153
  axiom 198
  equivalence 200
  flat problem 164
  future of 196
  game of 200
  how to count 180
  inconsistency defense 207
  inconsistency in 204
  indirect proof 244
  infinity 179
  linear problem 164
  long multiplication 240
  Lucas argument 205
  mathematical creativity 309
  non-deterministic polynomial problems 165
  Peano axioms 199, 205
  prime numbers 243
  PSPACE problem 168
  traveling salesman problem 166
  truth and rules 193, 200, 206
Matiyasevich, Yuri 248, 251
Mattapoisett 89
Maude, Isabel 14
Maxwell 298
Mayan astronomers 70
maze 164, 165
McChrystal, General Stanley A 54
McGinn, Colin 324
McGurk Effect 114
McLaughlin, Dan 37
Mehrabian, Albert 84
memory 11, 14, 90, 118
  digital 130
  photographic memory 119
  visio-spacial 28
  with age 120
memory management 250
meninges 97
meningitis 14
Merchant of Death 366
Merilees, Philip 170
Mesopotamia 87
metre 28
micro-expression analysis 84
microphones 289
Microsoft 237
Microsoft Word 135
microtubules 124, 281
Miles, Andrew 368
Millennium Falcon 348
Milliken 394
Minds, Machines and Gödel 205
Minessale, Anthony 265
Ming Dynasty 15
mirror neurons 116
Mirzakhani, Maryam 368
Mitra, Sugata 35
monkey 129
  moon shot story 144
monkeys and typewriters 254
Monty Python 93
Moon Base 348
Moore, Gordon 18
Moore's Law 18
Morgan, Edwin 138, 142
Morse code 336
motor cortex 31
Mozart 246, 298
MRI scan 69, 104
multiplication
  long 240
Munduruku tribe 181
Murphy's Law 226
muscle memory 90
muscles 117, 192
music 113
  computers and 258
musical compositions 8
myelin 31, 118

N

Napoleonic wars 129
NASA 59
Native American 89
Navier Stokes Hypothesis 238
nebula 174
Negroponte, Nicolas 35
neural network 116, 121, 280

neurons 50, 98, 116
  microtubules 281
  nerve impulse speed 99
  recovery time 121
  synapse 119
Newcastle University 35
Newman, Max 221
Newton, Isaac 171, 319
Newton's Rings 321
New York University School of Medicine 12
Niels Bohr Institute 327
nitrocellulose 218
Nixon, Richard 194
Nobel, Alfred 365
Nobel Prize 30, 50, 100, 102, 119, 213, 326, 365
non-algorithmic xi
non-computable processes 373
non-computable thought 261, 282
noncomputational creativity xi
non-determinism 275, 282, 318
non-deterministic behavior 44
non-deterministic polynomial problems 165
noninvasive imaging techniques 11
nonverbal communication 84
Norvig, Peter 57
No Silver Bullet 237
No Silver Bullet – Essence and Accidents of Software Engineering 231
Nova Institute 50
Nova Southeastern University 124
Nuclear Magnetic Resonance 104
numbers
  counting system 181
  defining 199
  Gödel numbers 203
  infinity 179
  nature of 155
  prime 156, 243, 343
  random 188
  real 186, 280
  Turing numbers 190
  zero 179
Nyquist, Harry xiii

O

Occam's Razor 68
omnipotence 340
omniscience 340
On Computable Numbers and their Application to the Entscheidungsproblem 211
One Laptop per Child 35, 82
one-time pad 216, 337
opium 55
optical illusions 113
oracle function 278
order-of-magnitude 164
O'Reilly, Edward 50
Organon 149
Orwell, George 178
Outliers 37
Oxford English Dictionary 86
Oxford University xi, 50, 149, 205, 282

P

paperclip test 298
paper tape 221
paradox xi, 68, 70, 114, 153, 203, 326
  Barber paradox 155
  EPR paradox 332
  false 155
  Kochen-Specker paradox 345
  liar's paradox 154
  Russell paradox 155
  Wiles paradox 247
  Zeno's paradox 154
parallel lines 179
Paramecium 123–126
particle accelerator 69
Pasteur Institute 188
pattern matching 49
Pavillon de Breteuil 28
PC revolution 236
PDP11-34 236
Peano axioms 199, 205
Peano, Giuseppe 197, 199
Pelmanism 28
pendulum 121
Penrose, Roger xi, 50, 56, 123, 205, 246, 261, 282, 309, 328
Penrose Steps 51, 113
Pentagon 79
people
  living forever 20
Pérez, Shimon 81, 82
permutation of information 135
Persia 15

personal computers 307
PET 107
philosophical proof 246
philosophy 193
Photo Electric Effect 366
photographic memory 119
photons 110, 320, 331, 332, 348
photosynthesis 50, 124
physics 180
  determinism 316
Picasso, Pablo 232, 246, 298
pit vipers 102
Pixar 306
Places game 168
plank interval 131, 255
Plato 149, 179
Podolsky, Jacob 332
poem
  computer created 134
Poincaré Conjecture 357
Poincaré, Henri 171, 180
Polaroid lenses 332
police strike 54
Polish Intelligence Bureau 211
polynomial 164
polywater 157
positivism 153
positron emission tomography 107
Post, Emil 258
Post Problems 258
Powell, Colin 53
PowerPoint 63
preparation 301
Previn, Andre 93
prime numbers 156, 243, 343
Princeton University 97, 212, 223, 251
Principia Mathematica 195
Private Eye 287
process
  creativity versus 312
program equivalence problem 250
programmer 235, 238, 263
  programming geniuses 271
  super-programmers 266
progressive cipher 214
proof
  Feynman's proof 157
  indirect 244
  mathematical 156
  philosophical 246
PSPACE problem 168
Pulitzer Prize 365, 367
purpose on the planet 9
puzzle
  two guards 151
Pygmalion 25
Pythagoras 243
Pythagorean triangle 240

Q

quantitative numerical skills 30
quants 71
quantum brains 122
quantum computers 277
quantum cryptography 215, 217
quantum effects 50, 282
quantum gravity interaction 329
quantum mechanics 22, 123, 180, 293, 344
  Copenhagen interpretation 327
quantum Morse machine 337
quantum randomness 45
quantum uncertainty 21
quartz 121
quartz crystal 43
qubits 21
quinine 100

R

radioactive decay 327
random chance 254
random numbers 44
Reagan, Ronald 79
reality 289
real numbers 280
reductio ad absurdum 152, 157
reflection 319
relativity 180
  Special Relativity 348
  Theory of Relativity 366
Renaissance 16, 297
Reykjavik 79
rice
  covering chessboard 163
Rice's Theorem 267
Riemann Hypothesis 238, 257
Riemann surfaces 368
Ritchie, Graeme 310
Rivest, Ron 165
Robinson Davis Matiyasevich theory 248

Robinson, Julia 248, 251
Robinson, Sir Ken 299
Rogers and Hammerstein 266
Rommel 211
Röntgen, Wilhelm 100
Rosen, Samuel 332
Rosetta Stone 89
Rowling, JK 306
Royal Swedish Academy of Sciences 366
RSA encryption 165
Rubik's cube 345
Rule 164 192
rules 192, 200
Rumsfeld, Donald 190, 194
RunMe 245
Ruohonen, Keijo 251
Russell, Bertrand 155, 193, 195
Russell paradox 155

S

Sachs, Oliver 12
Sagan, Carl 68
Sandler, Adam 12
Sanford, Edward 38
Schadenfreude 86
Scherbius, Arthur 211
Schrödinger, Erwin 326
Schrödinger's cat 325
Scientific American 85
scripts and symbols 87
Searle's Chinese Room 66, 222
secret message 214
Seeger, Pete 162
Sego, Daniel 335
self-halting problem 250
SendMail 265
Shadows of the Mind 50, 282
Shakespeare, William 39, 129, 131, 138, 140, 254
Shamir, Adi 165
Shannon, Claude xiii, 3, 216, 337
Sharman, Mike ix
Shaw, George Bernard 25, 78, 286
Shockley, William 30
Shor's algorithm 277
Siegelmann, Hava 280
silicon chip 20
Silicon Graphics 45
Simonsen, Inge 192
Singer, Isaac 314
Singh, Simon 242
single cell organisms 18
singularity 261
Siri 56
Skinner, BF 34
smallpox virus 189
smile 84
sock analogy 332
software 224, 231
  bugs 262
  coding 264
  drag and drop 237
  flowchart 244
  modern word processor 236
  origins of 238
  PC revolution 236
  problems 250
  process 270
  programmer 235, 238, 263
  programming geniuses 271
  re-architect 270
  scope creep 270
  spreadsheet 237
  super-programmers 266
  writing 235
sorting 164
Soviets 79
space like separation 348
Space Shuttle Columbia 58
Spanish 90
spears and hunters 181
Special Purpose Objection 253
Special Relativity 348
speed of light 317
spell checkers 139
spin 343
Spolsky, Joel 256
  Law of Leaky Abstractions 256
Spolsky's Law 267
spreadsheet 237
Stanford University 30
starch 118
Star Trek 157, 348
Star Wars 79
static intelligence 32
statistical approach 8
Stern, William 27
Stockhausen 7
Stoic logic 150
story 92

Strategic Defense Initiative 79
stroke damage 12
substitution code 203
Sumerians 87
superconductivity 22
super-programmers 266
super-Turing 275, 278, 283, 294
syllogisms 150, 152, 158
symbolic communication 87, 293
symbols 130, 143, 194, 201
synapse 119
system argument 66

T

Tagg, James
  Home Page 232
Tallis, Thomas 259
tapetum lucidum 110
TCP 267
Terman, Lewis 30
The Art of Thought 301
The Boy Who Can't Forget 120
The Cognitive Style of PowerPoint 63
The Emperor's New Mind xi, 50, 246, 282
The Free Will Theorem 343
The Free Will Universe 231
The Game of Logic 150
The God Delusion 339
The History of Western Philosophy 193
The Innovators Dilemma 306
The Journal of Irreproducible Results 139
The Labyrinth 150
The Man who Mistook his Wife for a Hat 12
The Mythical Man Month 231
Theory of Relativity 366
Thermal Imaging 101
The Sound of Music 266
The Talent Code 37
The Universe in a Nutshell 339
The Webby Awards 365
thiamin molecules 124
thinking 117
Thomas, Dylan 140
thought art 87
Three Body Problem 171
time intelligence 29
time travel 316
Tokyo University 13
Tolstoy, Leo 129
  War and Peace 129
toothed gears 43
Top Gear 97
topology 249
tortoise and hare 154
Torvalds, Linus 232, 265
totality problem 250
Tower of London 42
Transmission Control Protocol 267
traveling salesman problem 166
triple drug therapy 158
truth 192
tubulin 98, 118, 122, 124
  single-celled organisms 123
tubulin microtubules 282
tubulin molecules 50
Tufte, Ed 53, 63, 92
Tufts University 102
Turing, Alan xi, 15, 17, 26, 71, 96–97, 141, 209, 210, 211, 238, 258, 259, 302, 342
  and Churchill 213
  and computer understanding 212
  and Enigma 212
  and The Decision Problem 212, 219, 238
  and World War II 213
  at Bletchley Park 212
  at Cambridge University 211
  at Princeton University 212
  birth 211
  death 213
  homosexuality 213
  Turing Award 213
Turing Award 367
Turing limit 155, 253, 260, 275, 278, 280
Turing machine 212, 221, 246, 251, 275, 278, 280, 353
  Lego 223
  universal Turing machine 222
Turing numbers 190
Turing test 67, 71, 72, 261
Turing theorem 351
Tuszynski, Jack 119
twin particle experiment 331
twins 31
two guards puzzle 151
Two Ronnies 86, 93

U

Ultimate Question of Life the Universe and Everything 238
ultraviolet light 100
Uncanny Valley 75
uncertainty 318
understanding xiii, 53
  meaning of 56
  of computers 71
Unicode 129, 203
United States Army's Ballistic Research Laboratory 223
Universal Turing Machine 246
universe
  anthropic principle 322
  complexity and chaos 173
  determined 207
  deterministic 175
  end of 140
University of Alberta 119
University of Arizona 119, 122, 282
University of Calgary 251
University of California 280
University of Maryland 33
University of Massachusetts 280
University of Montreal 117
University of Moscow 258
University of Otago 33
University of Santa Cruz 7
University of Vienna 202
urban legend 11
US Supreme Court 28

V

van Gogh, Vincent 142
variable initialization 250
vertical abstraction 267
Visco, Frank L 396
vision 109
vision intelligence 29
visual agnosia 12
Volkswagen Polo 312
von Neumann, John 223
Vorderman, Carol 30

W

Wallas, Graham 1, 301
Wang 308
War and Peace 129
Watson computer 8, 207
Watson Research Laboratory 5
wave interference 320
wavelength 102
weather
  predicting 168, 172
Wechsler Test 27
Weizenbaum, Joseph 25
West, Mae 78
Whitehead, Alfred North 195
Who Wants to be a Millionaire? 347
Wikipedia 8, 129, 231
Wilder, Billy 128
Wiles, Andrew xi, 75, 242, 248, 251, 254, 259, 261, 351
Wiles Paradox 247
William of Occam 68
Winchester Drives 306
Wise, Michael 335
Wolfram, Stephen 173, 246, 352
Woods, Tiger 116
WordPress 237
World War I (First World War) 211, 337
World War II (Second World War) 17, 218, 337
  code breakers 211
Wozniak, Steve 10, 68
Wright, Stephen 274
WYSIWYG 237

X

XPRIZE 238, 367
X-ray 69, 100
  damage from 104
  slicing technique 102
  wavelength 102

Y

Yellow Pages 308

Z

Zar, Jerrold H 139
Zeitgeist 86
Zeno machine 280
Zeno's paradox 154
Zermelo-Fraenkel set theory 156
zigzag method 182

GEMs Paper #26

Saudi Arabia: beyond oil but not so fast

Merrill Lynch

30 June 2016 Corrected

Transforming Saudi Arabia, but challenges abound

The Saudi National Transformation Plan (NTP) bodes well for comprehensive reform efforts to diversify the economy in line with Vision 2030. The NTP a) identifies key sectors with high growth potential; b) starts to articulate supportive public sector industrial strategies; and c) seeks to foster higher value-added through enhancements to processes, products and organizations. Still, ambitious targets in a number of sectors are unlikely to be reached and medium-term macro sustainability is not yet clear.

USD peg holds but NTP inconsistent with stable FX policy

Our view is still that the USD peg holds. However, the NTP provides a mixed message. On the one hand, it blurs policymaking incentives given the need for a competitive Fx if diversification progresses. On the other, unrealistic fiscal targets mean consolidation is likely to fall short of easing imbalances materially without oil price recovery. Energy policy thus likely needs to become less aggressive to support macro and FX policies.

Eurobond premium required for fiscal slippage risk

We expect large and regular sovereign Eurobond issuance to support FX reserves and domestic liquidity but to weigh on regional bond spreads if risk appetite does not hold up or fiscal balance slips. EMBIG index inclusion is unlikely, in our view. Saudi forwards, CDS and rates are likely to stay under pressure due to issuance and tight liquidity.

Commodities: NTP adds to medium-term oil tightness

The NTP suggestion that production capacity is maintained until 2020 reinforces our conviction of medium-term oil market tightness. The NTP gas production target that could displace domestic crude demand and boost export capacity is challenging.

Equity Strategy: six investible themes on the back of NTP

We see a number of investible themes emerging from the NTP which investors can use to identify potential beneficiaries. These themes include rising availability of affordable housing, increased access to healthcare, down-trading as the consumer comes under pressure, a surge in telecom infrastructure and significant opportunities in the downstream petrochemical arena. Buy-rated names with access to these themes include: **Al Hammadi**, **Savola, Al Othaim**, **STC, Zain KSA** and **SABIC**.

Partial NTP success & MSCI inclusion could drive rerating

The Saudi market trades on a 12m fwd P/E of c.13x, an 11% discount to the long-term average and its lowest premium to GEMs since 2010. We believe that with earnings momentum gradually reaching an inflection point a relatively negative outlook is already being priced in. Given our view that oil prices will gain strong momentum in 2017, even a partial success of the NTP along with CMA market reforms to expedite inclusion in the MSCI EM index could be sufficient to improve confidence and thus drive a rerating.

Trading ideas and investment strategies discussed herein may give rise to significant risk and are not suitable for all investors. Investors should have experience in FX markets and the financial resources to absorb any losses arising from applying these ideas or strategies.

>> Employed by a non-US affiliate of MLPF&S and is not registered/qualified as a research analyst under the FINRA rules.

Refer to "Other Important Disclosures" for information on certain BofA Merrill Lynch entities that take responsibility for this report in particular jurisdictions.

BofA Merrill Lynch does and seeks to do business with issuers covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision.

Refer to important disclosures on page 82 to 85. Analyst Certification on page 81. Price Objective Basis/Risk on page 77.

GEM Fixed Income Strategy & Economics
Global

Jean-Michel Saliba
MENA Economist
MLI (UK)
+44 20 7995 8568
jean-michel.saliba@baml.com

Hootan Yazhari, CFA >>
Research Analyst
Merrill Lynch (DIFC)
+971 4 4258218
hootan.yazhari@baml.com

Francisco Blanch
Commodity & Deriv Strategist
MLPF&S
+1 646 855 6212
francisco.blanch@baml.com

Faisal AlAzmeh, CFA >>
Research Analyst
Merrill Lynch KSA Company
+966 11 299 3741
faisal.alazmeh@baml.com

Abdelrali El Jattari >>
Research Analyst
Merrill Lynch (DIFC)
+971 4 4258231
abdelrali.eljattari@baml.com

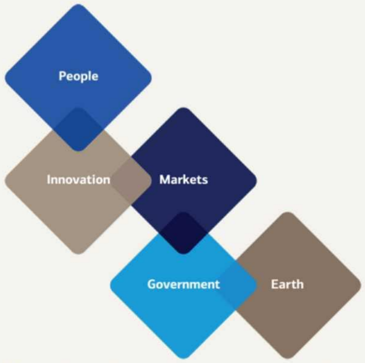
Jamie Clark, CFA >>
Research Analyst
MLI (UK)
+44 20 7995 1300
jamie.clark@baml.com

Ali Dhaloomal
Research Analyst
MLI (UK)
+44 20 7996 9107
ali.dhaloomal@baml.com

Celine Fornaro >>
Research Analyst
MLI (UK)
+44 20 7996 5515
celine.fornaro@baml.com

See Team Page for Full List of Contributors

A Transforming World



This report ties into the enterprise-wide Investment Themes recently introduced in A Transforming World. In particular, the Saudi National Transformation Plan is likely to span the themes of Markets (Frontier), Government (Reform) and Earth (Energy Efficiency)

Unauthorized redistribution of this report is prohibited. This report is intended for amanda.ens@baml.com

Contents

Macro: charting the way forward 3

Towards better governance 3

Public Investment Fund to gain prominence 7

Historic energy sector liberalization 10

National Transformation Plan: many promises, few details 13

Lessons in diversification 31

Eurobond premium required for fiscal slippage risk 34

Commodities: NTP adds to medium-term oil market tightness 36

Equity Strategy: more clarity required, but investible themes emerging 39

Six key investible themes from the NTP are emerging 40

Saudi market valuation not testing 42

Taking positive steps to accelerate MSCI inclusion 45

A higher weighting in international indices 45

Telecom: supporting a move to higher connectivity 48

NTP helps increase penetration of higher margin offerings 48

Health: Not Tremendously Prescriptive 51

Vision 2030 a start but Needs Transparent Proposals 51

Healthcare implications of National Transformation Plan 51

Consumer: a necessary pain 58

Prefer staples over discretionary 58

Rationalization of subsidies for water and electricity 60

Saudi Arabia approves 100% foreign ownership rules 61

Real estate: NTP positive but not enough 64

Key strategic objectives and KPIs to watch 64

Land tax - gradual revenue source 65

Metals & Mining: ambitious growth target 67

Government focus on mining to boost jobs and growth 67

Oil & gas/petchems: focus on downstream 69

Focus on downstream expansion is a priority 69

Refining capacity target: almost there 71

Ambitious natural gas expansion 72

Utility sector: sharing the capex burden and achieving cost reflective tariffs 73

Eight strategic objectives aimed at being self-funding 73

Defence: Vision 2030 supports defence spending 75

Saudi Arabia to build out military industrial base 75

Research Analysts 86

Macro: charting the way forward

Jean-Michel Saliba
MLI (UK)
jean-michel.saliba@baml.com

The Saudi Vision 2030 and National Transformation Plan (NTP) present a comprehensive roadmap for change and augur for committed diversification efforts, in our view. Announced reforms can help sustain higher potential growth if fully realized as new sectoral sources of growth are developed. Positively, the NTP borrows elements from other successful strategic case studies. Still, we find a number of targets ambitious or unrealistic on the fiscal and diversification fronts, and the sequencing and details of the fiscal measures are left nebulous.

In terms of the other pillars of the macro view, namely energy and Fx policy, the NTP presents a mixed picture, in our view. Energy policy is likely to be less aggressive, as the NTP suggests a constant oil production capacity. To remain consistent with the targeted government debt accumulation path, we estimate that oil prices have to average at least US$50/bbl in 2016-20 along with no growth in spending (excluding the additional cost of NTP initiatives). Partial implementation could require oil prices of cUS$65/bbl.

Fiscal consolidation remains imperative to support the Fx peg, in line with stability and economic imperatives. However, as diversification progresses, the case for increased Fx flexibility to support competitiveness could likely gradually take shape. Policy-making growth-focused plans may conflict with needs to deflate Fx demand in the economy.

Towards better governance

The comprehensive economic blueprint unveiled by the Saudi Vision 2030 and the National Transformation Plan (NTP) confirms the economic reform credentials of the current administration. It is likely to improve government culture, accountability and transparency, in our view. The NTP will be implemented across 24 government bodies and has been presented through a number of press conferences with high-level ministerial presentations. The Governance Framework details steps to institutionalize and coordinate implementation through restructuring the government. It proposes a number of committees and bodies to report to the Council for Economic and Development Affairs (CEDA). It also introduces an escalation mechanism to rapidly resolve bottlenecks, which could see matters escalated to CEDA in 42 days.

A wide-ranging policy-making reshuffle

In line with ongoing Saudi Vision 2030 implementation efforts, a wide-ranging restructuring of the Saudi Cabinet and government bodies was announced in May to drive change in the government. King Salman issued 51 Royal Decrees restructuring the Cabinet, various government bodies and appointing a number of officials into various government roles. A statement concurrently issued by the Royal Court suggested these changes are in line with the recently announced Saudi Vision 2030 and aim to support its implementation. This follows from the January 2015 restructuring of the bodies affiliated to the Council of Ministers, which saw the creation of CEDA and the Council for Political and Security Affairs. We list a few major changes below.

New Oil Minister represents technocratic experience and policy continuity

The appointment of Khalid Al-Falih to the Ministry of Energy, Industry and Mineral Resources is not surprising, given that former Minister of Petroleum and Mineral Resources Al-Naimi has previously suggested he was approaching retirement due to old age. Al-Falih's appointment continues the tradition of non-Royals heading the Ministry, while simultaneously holding the Chairmanship role of Saudi Aramco. Furthermore, according to local press, Al-Falih appears close to Deputy Crown Prince Mohamed bin Salman, and his appointment likely confirms the recently assertive Royal influence over energy policy, in our view. New Minister of Energy, Industry and Mineral Resources Al-Falih's past official pronouncements are consistent with stable Saudi energy policy, in our view.

Expanded role for the Oil Ministry under Vision 2030

The Ministry of Petroleum and Mineral Resources has been expanded to become the Ministry of Energy, Industry and Mineral Resources. It will be dedicated for energy, in addition to expanding to encompass responsibilities relating to electricity and industry. It will also undertake the management of the National Industrial Cluster Development Program (NICDP). The new Minister will chair the board of directors of the Royal Commission for Jubail and Yanbu, the Industrial Development Fund, the Saudi Organization for Industrial Estates and Technology Zones, the Saudi Geological Survey, the King Abdulaziz City for Science and Technology, the Saudi Exports Development Authority and the King Abdullah City for Atomic and Renewable Energy.

Energy sector liberalization drives non-oil diversification

The wider role for the Ministry of Energy, Industry and Mineral Resources is in line with the expanded industrial responsibilities foreseen for Saudi Aramco in the Vision 2030. It may also suggest a continued desire to liberalize the energy sector. The National Industrial Cluster Development Program (NICDP), established in October 2012 by the Ministry of Petroleum and Mineral Resources and the Ministry of Commerce and Industry, could likely gain further importance in the industrialization strategy adopted in the Vision 2030. The NICDP will be managed by the Ministry of Energy. It seeks to leverage Saudi Arabia's comparative advantage in energy, petrochemicals and minerals to create sustainable export oriented industries, initially focused on the automotive, construction material, appliances, metal processing and flexible packaging sectors.

Restructuring the Cabinet and other government bodies

The Ministry of Commerce and Industry will become the Ministry of Commerce and Investment and sees the appointment of a new Minister. The new Minister will chair the board of directors of the General Authority for Investment, the General Authority for Small and Medium Enterprises and the Saudi Standards, Metrology and Quality Organization. The Ministry of Labor will be merged with the Ministry of Social Affairs to form the Ministry of Labor and Social Development, and sees the appointment of a new Minister. The Ministry of Water and Electricity has been disbanded. The Ministry of Agriculture was renamed the Ministry of Environment, Water and Agriculture and expands its responsibilities into these areas. Two new Ministers for Transport and Health were appointed. The Ministry of Hajj has been renamed the Ministry of Hajj and Umrah and sees the appointment of a new Minister.

The Saudi Fund for Development will now report to the Council for Economic and Development Affairs. The General Authority for Entertainment and the General Authority for Culture have been created. The Department of Zakat and Income Tax was renamed the General Authority for Zakat and Income, and will report to the Ministry of Finance.

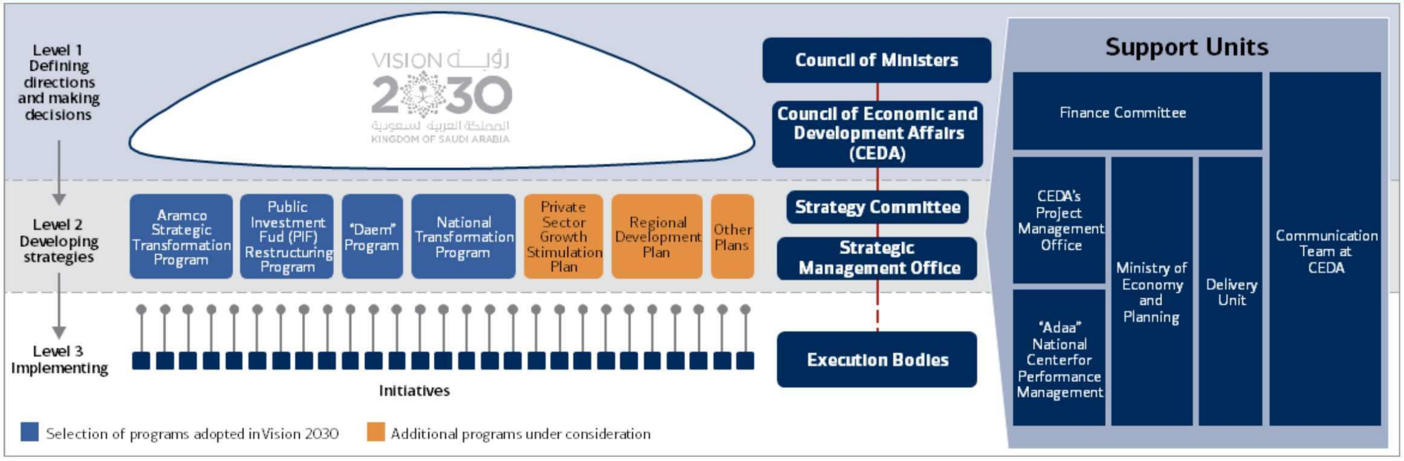
New Central Bank Governor maintains commitment to USD peg

Saudi Arabian Monetary Agency (SAMA) Governor Al-Mubarak has been replaced in his post by SAMA Deputy Governor for Research and International Affairs Dr. Al-Kholifey. A number of officials were also appointed into various government and advisory roles. New SAMA Governor Al-Kholifey's past policy pronouncements suggest continued commitment to the USD peg.

Watch the Saudi Fund for Development

The Saudi Fund for Development (SFD) new direct link to the Council for Economic and Development Affairs likely concentrates further authority in the Deputy Crown, in our view. Given that the Public Investment Fund (PIF) last saw similar links being established prior to plans being made for it to be turned into a Sovereign Wealth Fund (SWF), this could suggest further strategic changes may take place at the SFD.

Exhibit 1: Saudi Vision 2030 Governance Model



Level 1 Defining directions and making decisions
VISION رؤيـــة 2030 المملكة العربية السعودية KINGDOM OF SAUDI ARABIA
Council of Ministers
Council of Economic and Development Affairs (CEDA)
Level 2 Developing strategies
Aramco Strategic Transformation Program
Public Investment Fund (PIF) Restructuring Program
"Daem" Program
National Transformation Program
Private Sector Growth Stimulation Plan
Regional Development Plan
Other Plans
Strategy Committee
Strategic Management Office
Level 3 Implementing
Initiatives
Execution Bodies
Support Units
Finance Committee
CEDA's Project Management Office
"Adaa" National Centerfor Performance Management
Ministry of Economy and Planning
Delivery Unit
Communication Team at CEDA
Selection of programs adopted in Vision 2030
Additional programs under consideration

Source: Saudi Vision 2030

A web of Strategic Programs

Saudi Vision 2030 incorporates a number of Executive Programs and initiatives, with implementation carried out by several government entities. The Vision continues to centralize decision-making into CEDA, chaired by Deputy Crown Prince Mohammed bin Salman, and reporting to the Council of Ministers. The Ministry of Economy and Planning also retains an important role, with lower emphasis being put on the Ministry of Finance. The Project Management Program also suggests emphasis on ongoing review of capital expenditures, whereby existing review served to examine their approval process, and control their level. From a macro perspective, the three most important programs for now are: a) the re-shaping of the Public Investment Fund (PIF) into a US$2trn Sovereign Wealth Fund (SWF); b) the new corporate strategy for Saudi Aramco to transform it into an energy and industrial conglomerate; and, c) the National Transformation Plan (NTP) which encompasses medium-term growth boosting initiatives and fiscal consolidation measures (alongside the Fiscal Balance and Privatization Programs). We discuss these three elements in turn below.

Table 1: Strategic Programs introduced by the Saudi Vision 2030

Strategic Program	Comment
Government Restructuring Program	Supreme Councils have already been implemented, and the Council of Political and Security Affairs and the Council of Economic and Development Affairs (CEDA) have been established
Strategic Directions Program	Strategic directions determined by state agencies and approved by the government
Regulations Review Program	Several laws have been reviewed or enacted already such as the Company Law, the Non-Governmental Organizations Law, the White Land Law, and the General Authority for Endowments (Awqaf) Law
Performance Measurement Program	Center for Performance Management of Government Agencies has been established
Human Capital Program	Aims to measure, assess, analyze and support the efficiency of civil service
Program for Strengthening Public Sector Governance	Strategic Management Office reporting to CEDA as well as a Decision Support Center at the Royal Court are to be established
Strategic Partnerships Program	Aims for stronger ties with economic partners to enhance exports
"Daem" Program	Aims to enhance the quality of cultural activities and entertainment
Private Sector Growth Stimulation Program	Under consideration
Regional Development Program	Under consideration
Fiscal Balance Program	Likely target of achieving fiscal balance by 2020
Project Management Program	Expert Project Management Offices (PMOs) and a Central Delivery Unit have been established
Saudi Aramco Strategic Transformation Program	Aims to position Saudi Aramco as a leader in more than one sector
Public Investment Fund (PIF) Restructuring Program	Aims to transform the PIF into the largest Sovereign Wealth Fund (SWF) in the world
Privatization Program	Comprehensive privatization program; targets under study
National Transformation Plan Program	Interim Key Performance Indicators (KPI) targets to achieve by 2020

Source: Saudi Vision 2030, BofA Merrill Lynch Global Research

Table 2: Initiatives introduced by the Saudi Vision 2030

Initiative	Current	Target
Religious tourism		
Number of yearly Umrah visitors per year (mn)	8	30
Building the largest Islamic museum in the world	-	-
Culture		
Number of Saudi heritage sites registered with UNESCO	4	8
Number of Saudi cities recognized in the top-ranked 100 cities in the world	0	3
Household spending on cultural and entertainment activities (%)	2.9	6
Social		
Average life expectancy (years)	74	80
Unemployment rate (%)	11.6	7
Female labour force participation (%)	22	30
Household savings ratio (% of household income)	6	10
Ranking in the Social Capital Index	26	10
Ratio of individuals exercising at least once a week (%)	13	40
Number of volunteers per year	11,000	1,000,000
Localization		
Localization of oil and gas sectors (%)	40	75
Localization of defence industry (%)	2	50
Renewable energy value chain (%)	-	-
Sovereign Wealth Fund		
Public Investment Fund assets (SAR bn)	600	7,000
Economy		
Global ranking of the economy in terms of size	19	15
Ranking in the Global Competitiveness Index	25	10
Ranking in the Logistics Performance Index	49	25
Private sector contribution to GDP (%)	40	65
Foreign Direct Investment (% of GDP)	3.8	5.7
Non-oil exports share in non-oil GDP (%)	16	50
SME contribution to GDP (%)	20	35
Non-profit sector contribution to GDP	<1	5
Improving the business environment and pursuing public-private partnerships	-	-
Rehabilitating economic cities and restructure King Abdullah Financial District	-	-
Establishing special zones such as logistic, tourist, industrial and financial ones	-	-
Revise energy subsidies and redirect support to eligible citizens and economic sectors	-	-
Ease restrictions on ownership and foreign investment in the retail sector	-	-
Fiscal		
Non-oil government revenue (SAR bn)	163	1,000
Governance		
Ranking in the Government Effectiveness Index	80	20
Ranking in the E-Government Survey Index	36	5

Source: Saudi Vision 2030

Public Investment Fund to gain prominence

The restructuring of the Public Investment Fund (PIF) is likely to allow greater focus on achieving a diversified foreign asset base which could support in turn the build-up of non-oil revenues. According to Deputy Crown Prince Mohammed bin Salman, government ownership of Saudi Aramco would be transferred to the PIF, which will be transformed into a SWF that will look to increase its overseas assets (from 5% of total to 50% of total by 2020) following its recapitalization and the proceeds of Aramco IPO. The PIF would hold on-paper a vast amount of wealth post-IPO (US$2trn, according to the Deputy Crown Prince, the bulk of which would be Saudi Aramco) as ownership of Saudi Aramco is transferred to the PIF, but it would only be able to deploy the cash proceeds of the monetized Aramco stake in the near-term, in our view. We think the PIF's transformation is still at a relatively early stage for now.

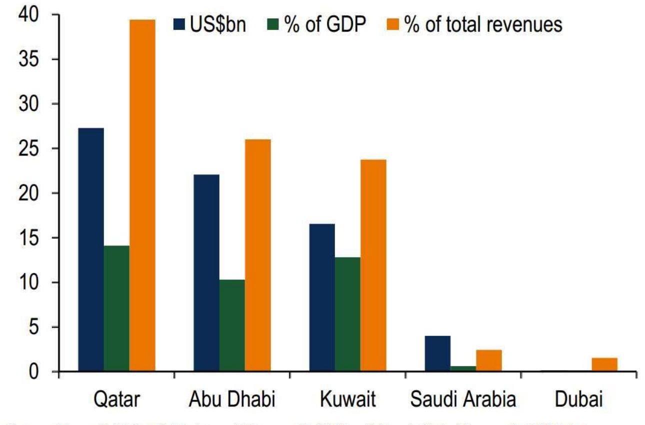
Room to grow PIF stature and budget contribution

The restructuring of the PIF is likely to allow further diversification of foreign assets, which will in turn increase the share of investment income in the budget over time. We estimate that in 2015 investment income transferred to the budget stood at US$9.9bn (1.5% of GDP), which entails transfers of US$4bn from the PIF and US$5bn from SAMA. These transfers pale in comparison to the estimated budgetary contributions among main GCC peers. We estimate that the rate of return (investment income) on foreign assets of SAMA, government entities and the private sector averaged c2% over the past 8 years, which already suggests some exposure to riskier asset classes, in our view. As the PIF gains importance, it will become more prominent in the examination of the breakdown of the Saudi Net International Investment Position .

Transitioning to an Abu Dhabi model

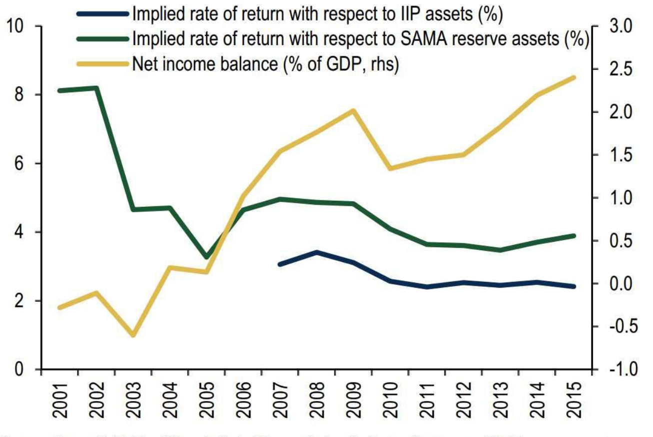
It will be interesting to see how the restructuring of the PIF into an SWF works out in practice. We hypothesise that it may be that, on top of the monetization of Aramco's stake sale, PIF could get a portion of the assets of SAMA. In this scenario, we would effectively transition to the Abu Dhabi and Kuwait model where the central bank holds little reserves and non-transparent SWFs are what matters both in terms of flow and stock. Foreign assets purchases of the PIF would also have to be managed within the overall Balance of Payments (BoP) framework as they could lead to drains on SAMA reserves in the near term. Over time, Saudi Arabia could decide to emulate the Norway model, which would entail a more prudent use and conduct of fiscal policy, in our view.

Chart 1: PIF budget contributions small versus GCC SWF contributions



Source: Haver, IMF, Saudi Ministry of Finance, BofA Merrill Lynch Global Research. 2015 data, Investment income and transfer of profits of public entities for Kuwait. Investment income from public enterprises (incudes Qatar Petroleum's net income) for Qatar.

Chart 2: High rate of return suggests foreign assets well diversified



Source: Haver, BofA Merrill Lynch Global Research. Implied rate of return on SAMA reserve assets simplistically assumes all investment income is earned by SAMA (instead of being earned by SAMA, government entities and the private sector).

PIF largely a domestic inward-looking entity until recently

The IMF reports that the PIF had assets of 11.1% of GDP in 2014 (c.SAR310bn). PIF has foreign assets of SAR14.1bn as of end-2015. It has outstanding loans of SAR103.9bn as of 3Q15 (excluding electricity loans of SAR14bn that PIF administers). According to Bloomberg, PIF had stakes in publicly listed companies on Tadawul worth SAR934bn as of April 2016. According to the MoF, PIF held equity worth SAR63.3bn in 41 Saudi companies in total at end-2011, as well as stakes worth SAR14.9bn in a number of pan-Arab corporations.

In July 2014, the cabinet authorized the PIF to establish companies inside and outside Saudi Arabia, alone or in partnership with other institutions from the public or private sectors. It also bought a US$1.1bn stake in a South Korean company and recently announced in early June it took a US$3.5bn stake in Uber, denying in the process that it was considering a US$3bn loan to fund the acquisition. PIF has been asked to co-invest US$10bn in Russia or the Middle East with the RIDF. Note that PIF has established Sanabil investments with SAR20bn in capital. The PIF has paid dividends to the budget last year (SAR15bn). It now reports directly to the Council of Economic and Development Affairs headed by the Deputy Crown Prince, after reporting to the Ministry of Finance in the past.

JASTA bill not a hurdle for further Saudi investment in the US

We expect that the deep and liquid US financial markets will remain a prominent destination for Saudi foreign asset holdings. The potential passage of the "9/11" bill (Justice Against Sponsors of Terrorism Act, JASTA) in US Congress raises the possibility that Saudi Arabia would choose to liquidate up to US$750bn in assets in the US, according to Saudi Foreign Minister Adel al-Jubeir. A forced and rapid liquidation of US-based Saudi foreign assets may expose SAMA and other government entities to mark-to-market losses, and require changes to Fx reserve management

The language of the JASTA bill as it currently stands does not suggest an imminent Saudi sell-off of US assets, in our view. The "Stay of Actions Pending State Negotiations" section inserted into the bill allows a stay to be granted for an indefinitely renewable 180-day period. This would be subject to court petition by the Attorney General and repeated certification by the Secretary of State that the US is "engaged in good faith discussions with the foreign state defendant concerning the resolution of the claims against the foreign state". The certification thus depends on the current and future US administration foreign policy inclinations, in our view. Separately, we note that the White House spokesman said US President Obama did not support the legislation and likely would not sign it. (A potential presidential veto may however still be overruled if Congress assembles the necessary two–thirds vote of each house). We continue to expect a close relationship between the US and Saudi Arabia given the confluence of interests in a broad range of matters.

Table 3: Breakdown of Saudi Arabia Net International Investment Position (US$bn)

	2007	2008	2009	2010	2011	2012	2013	2014	2015
Net International Investment Position	**380**	**471**	**435**	**479**	**585**	**685**	**763**	**792**	**703**
% of GDP	91	91	101	91	87	93	103	105	108
Net Foreign assets (excluding SAMA)	74	28	24	34	41	28	37	59	87
% of GDP	18	5	6	6	6	4	5	8	13
Assets	**495**	**631**	**636**	**708**	**824**	**936**	**1,028**	**1,069**	**993**
Non-reserve assets	190	188	226	262	280	279	302	337	376
by holder:									
Investment funds	7	4	4	5	5	5	6	8	7
Commercial banks	39	41	56	52	56	57	56	67	84
Public Investment Fund (PIF) foreign investment	1	2	2	4	4	5	4	4	4
Saudi Fund for Development (SFD) cumulative loan disbursements	6	6	6	7	7	7	8	8	-
Public Pension Agency (PPA) and General Organization for Social Insurance (GOSI) net foreign assets	-	49	-	-	-	98	112	-	-
of which, managed by SAMA (e)	43	48	51	59	65	69	75	80	78
PPA net foreign assets	-	24	-	-	-	63	72	-	-
GOSI net foreign assets	-	25	-	-	-	35	40	-	-
Development funds and other government (exc. PPA/GOSI) entities' foreign assets managed by SAMA (e)	15	17	18	20	22	23	25	27	26
Other	-	69	-	-	-	85	92	-	-
by instrument:									
Direct investment abroad	17	20	23	27	30	34	39	45	63
Portfolio investment	105	99	122	146	158	168	178	199	202
Equity securities	53	50	59	79	82	92	99	111	109
Debt securities	52	48	63	68	77	75	79	88	94
Other investment	68	69	82	90	92	77	85	93	111
Loans	3	4	4	3	3	2	2	1	1
Currency and deposits	55	63	73	76	78	69	76	85	105
Other assets	10	3	5	11	11	6	7	6	5
SAMA Reserve assets	306	443	410	445	544	657	726	732	616
Currency and deposits	93	132	111	117	148	197	191	187	204
Foreign securities	211	308	286	315	380	444	519	532	400
Other assets	2	3	13	13	16	16	15	13	12
Liabilities	**-116**	**-160**	**-201**	**-228**	**-239**	**-251**	**-265**	**-278**	**-289**
Direct investment in reporting economy	-73	-113	-148	-176	-187	-199	-208	-216	-224
Portfolio investment	0	-3	-4	-4	-6	-10	-17	-17	-17
Equity securities	0	0	0	0	0	-9	-15	-15	-15
Debt securities	0	0	0	0	0	-1	-2	-2	-2
Other investment	-42	-44	-49	-48	-46	-42	-40	-45	-48
Loans	-13	-14	-13	-13	-12	-11	-9	-10	-12
Currency and deposits	-28	-28	-25	-25	-23	-20	-19	-25	-26
Other assets	-1	-1	-11	-10	-11	-11	-11	-11	-11
memo:									
Identified Saudi Arabian and Middle Eastern holdings in the US									
Saudi Arabia US Treasury (UST) holdings	51	46	83	69	68	76	80	91	109
Long-term securities	50	45	82	68	67	74	78	88	107
Short-term securities	1	1	1	1	1	2	2	3	2
Saudi Arabia holdings of US equities	45	42	31	36	53	65	68	78	52
Saudi Arabia holdings of US Agency debt securities	8	42	14	7	5	5	7	6	6
Long-term securities	8	41	14	7	5	5	7	6	6
Short-term securities	0	1	0	0	0	0	0	0	0
Saudi Arabia holdings of US corporate debt securities	7	14	17	12	12	9	15	19	15
Long-term securities	5	9	11	9	8	6	10	13	13
Short-term securities	2	5	6	3	4	3	5	5	3
Middle East Foreign Direct Investment (FDI) position in the US (cost basis)	9	9	10	8	10	11	12	11	-
Middle East oil exporters gross claims on US banks	81	122	106	101	136	132	134	129	120
Saudi Arabia deposits in Bank for International Settlements (BIS) reporting banks									
Cross-border deposits with BIS reporting banks	163	180	169	167	188	244	224	210	212
of which, implied deposits from banks and others	102	142	128	113	135	180	174	158	170
of which, deposits from nonbanks	60	39	41	54	53	64	50	52	42
Saudi Fund for Development									
Cumulative signed loans	7	8	8	9	10	11	11	12	-
Cumulative disbursements	6	6	6	7	7	7	8	8	-
Government entities foreign assets managed by SAMA off-balance sheet	*58*	*66*	*69*	*79*	*87*	*92*	*100*	*107*	*104*
Deposits with banks abroad	4	5	3	2	2	6	13	10	5
Foreign securities	55	61	66	77	85	85	87	97	98

Source: Haver, SAMA, IMF, SFD, BofA Merrill Lynch Global Research. Middle East oil exporters consist of Bahrain, Iran, Iraq, Kuwait, Oman, Qatar, Saudi Arabia and the United Arab Emirates.

Historic energy sector liberalization

Partial privatization of energy sector assets could represent one of the cornerstones of the ambitious medium-term diversification strategy, and, if confirmed, serve as a highly visible and historic milestone. Such privatizations could ease key macro concerns regarding unsustainable debt accumulation and Fx reserves drawdown. We believe it would also confirm the government's macro reform credentials, its preparedness for a prolonged period of low oil prices and likely signal no near-term energy policy capitulation. The NTP suggestion of a flat production capacity through 2020 suggests less aggressive market share strategy, in our view.

Privatization of energy sector could start by 2018

In a recent interview with Bloomberg, Saudi Deputy Crown Prince Mohammed bin Salman outlined an array of measures the Saudi government could introduce to help diversify the economy, reduce its reliance on the hydrocarbon industry and ultimately support medium-term fiscal consolidation in a low oil price environment. Amongst these measures, the Deputy Crown Prince included a potential Initial Public Offering (IPO) of Saudi Aramco (at the parent level, and including various subsidiaries). With close to 250bn bbls of oil reserves, more than 10mn bbls per day of production and the world's 4th largest gas reserves, Saudi Aramco is the leader in hydrocarbon access among state owned, private and public Oil & Gas (O&G) companies alike. Saudi Aramco is the world's largest Oil & Gas Company by production and arguably the Kingdom's most important asset. Specifically, he outlined a plan that would see "less than 5%" of the company being sold to investors on the Saudi stock exchange no later than 2018 (but potentially as soon as 2017). Furthermore, he indicated the Kingdom is seeking to transform Aramco into an "energy-industrial company". The high profile stake sale would provide a highly visible anchor to diversification efforts and capitalize its newly restructured SWF.

Energy assets form a major and strategic sector for the economy

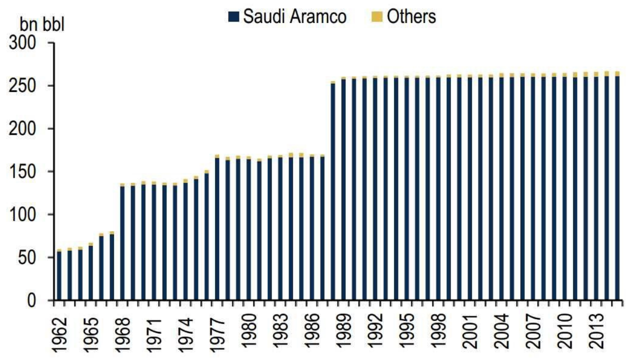
It is hard not to understate the central role the energy sector plays in the Saudi economy. The hydrocarbon sector represented 43.0% of the economy in real terms in 2015. Deputy Crown Prince Mohammed bin Salman suggested a US$2trn price tag for Saudi Aramco in his Bloomberg interview. Aramco is deeply intertwined with the Saudi state, and its workforce Saudization has increased over the years to 83%. Furthermore, it is the foundation of its state's domestic and international energy policy. As such, a decision to partially privatize the energy sector will be a momentous one to be carefully examined by the technocratic and political leadership, in our view.

More questions than answers at this stage

Deputy Crown Prince Mohammed bin Salman's initial suggestion in an interview with The Economist that a Saudi Aramco IPO could be considered has raised market expectations. Ensuing comments from Aramco's management did not make clear whether a listing would occur, and if so, in what form (upstream versus downstream), although the Deputy Crown Prince's subsequent Bloomberg interview suggested further impetus being given to studying the matter.

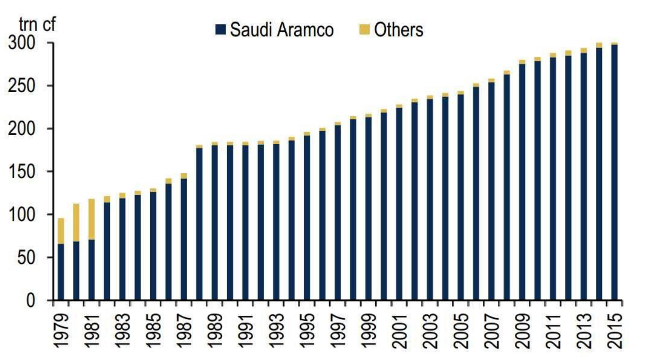
A listing of the entity itself may force much more transparency regarding hydrocarbon reserves (broadly unchanged for several years), lifting costs (local crude oil sales suggest cUS$5/bbl if no loss is incurred) and the fiscal regime (estimates in the press have suggested that Aramco's profits are taxed at 85%, with royalties set at 20%, but we also note from fiscal data that the fiscal transfer ratio in oil revenues to the Ministry of Finance appears to drop in years of low oil prices). It may also conflict with non-commercial ventures or strategic aspects of Saudi Aramco. This would include the maintaining of spare capacity and state policy matters (including energy policy). Note for instance that the US is the second regional destination for Saudi crude oil exports (17.5% of total Saudi crude oil exports) behind Asia. This was due to a strategic geopolitical decision to remain an important crude oil supplier to the US and was achieved through the buying of stakes in refineries in 1988.

Chart 3: Saudi Arabia proven crude oil reserves



Source: SAMA, Bank of America Merrill Lynch Global Research

Chart 4: Saudi Arabia proven gas reserves



Source: SAMA, Bank of America Merrill Lynch Global Research

Oil sector privatization could slow debt build-up and support Fx reserves

Potential energy sector asset partial privatization could raise a substantial amount for the Saudi government, in our view. A share sale would be a non-debt creating financing flow for the budget, which would lessen the need to borrow domestically or externally. A share sale could also increase Saudi Arabia's ability to project soft power and influence regionally and internationally. However, it would only support the build-up of Fx reserves in the case of foreign investor participation. Assuming a US$2trn valuation suggested by the Deputy Crown Prince as the value for the SWF, the vast majority of which would be Aramco through a possible 5% listing, such a listing could raise up cUS$100bn, which would represent the current annual government borrowing requirements and associated Fx reserves drawdown. We expect further administered energy price adjustments over the next five years, which should increase state revenues, in our view.

A share sale may also facilitate international borrowing on the back of Saudi Aramco's balance sheet. Saudi Aramco also announced it would look to set up a corporate debt program to fund its NTP initiatives, although issuance does not appear to be an immediate prospect. A corporate debt program, alongside a sovereign one, could allow the government to simultaneously tap a new pool of liquidity through corporate bond investors, upstream proceeds if needed and support Fx reserves. This could be achieved without impacting central government debt ratios and without saturating the government debt issuance room with sovereign debt investors. Saudi Aramco raised a US$10bn syndicated loan in late March 2015 in part to fund acquisitions. The loan replaced US$4bn in existing facilities agreed to in 2010, and includes a US$6bn five-year tranche with two 1-year extension options, a US$1bn 1-year renewable facility, a non-interest bearing SAR7.5bn (US$2bn) five-year tranche with two 1-year extension options and a SAR3.75bn (US$1bn) 1-year tranche renewable yearly.

Very roughly, taking the US$2trn Deputy Crown Prince potential valuation for Aramco, and assuming an optimal 10% net-debt-to-equity medium-term target, it would suggest that Aramco could manage leverage of up to US$200bn (c30% of GDP) in domestic and external debt on its balance sheet. Within a corporate setting, possible higher gearing needs to be justified to shareholders and could reflect financing of acquisitions or NTP costs. From a macro angle, international issuance would support Fx reserves.

Three aspects of energy policy to consider going forward

Energy sector privatization is likely to raise questions on the future path of energy policy, though we believe the latter will stay the course for now. We believe that are three aspects that are central to energy policy in Saudi Arabia: a) competition for global market share in upstream operations; b) future investment decisions and maintaining of spare capacity buffer; and c) integration and expansion of downstream ventures to create domestic jobs, secure captive crude demand and make operations less volatile.

Targeting a stable oil price

As a matter of Saudi policy, we would expect a preference for oil market stability over volatility, in line with recent official pronouncements. Although volatility makes it more difficult for high cost unconventional producers to operate, volatility is also damaging for Saudi Arabia and long-term planning domestically. For instance, no budget was announced for 1986 due to uncertainties in the world oil market, with monthly current appropriations set at one-twelfth of estimated actual expenditures for the prior year. We think Saudi authorities view possible eventual sharp upward price spikes in the oil price as damaging to both oil-consuming and oil-producing countries.

Future investment decisions will be critical for oil prices

Spare capacity is an important policy parameter for Saudi Arabia. In addition to its influence on prices through its ability to adjust production, Saudi can decide on the pace of its reserves development which would potentially affect future supply. There are distinct trade-offs in this decision. If Saudi Arabia is producing at close to its maximum capacity, it will have little control over sharp upside movements in oil prices. If excess capacity is large, oil prices are likely to be under downward pressure.

In the near-term, given the uncertainty in the market, in our view, there is likely little incentive for Saudi Arabia to embark on a program to invest and increase capacity from the current level. New investments may still be made to offset declines in existing fields. Over time however, our commodities research medium-term oil balances suggest that Saudi's spare capacity would be eroded over time unless further increases in spare capacity take place. Our BofAML commodities research medium-term supply-demand balances suggest an increased call on OPEC and need to increase production capacity. In 2004-09, Aramco invested cUS$100bn into its largest-ever capacity expansion program, which increased total capacity from 10mn bpd to 12.5mn bpd. For now, Aramco has suggested it will not slow down its hydrocarbon capex program for the next three years as cuts were achieving through savings in drilling costs and supplier discounts.

Energy policy is likely to be less aggressive going forward

We expect a less aggressive energy policy going forward, particularly as the NTP suggests a constant oil production capacity and an increase in gas production domestically. This is in line with Saudi Minister of Energy, Industry and Mineral Resources Al-Falih's pronouncements during the June Ordinary OPEC meeting. While his comments served to project increased OPEC institutional credibility and increased rapprochement with fellow OPEC members, he suggested that Saudi Arabia will remain responsive to customer needs but dismissed fears that increasing Saudi market share would take place in an aggressive manner (that could drive prices lower again).

Downstream operations - increased focus

Saudi Aramco's downstream integration drive is likely being pursued for two core reasons: (1) to support demand for Saudi crude (placing own crude in own refining capacity) and (2) a risk reduction mechanism in a highly volatile oil environment (having an integrated value chain allows capture of margins from well to the forecourt, promoting stability). In reflection, Saudi Aramco is building highly sophisticated refineries domestically, in addition to already completed joint-ventures and stakes in refineries abroad (China, US, Japan, South Korea, Indonesia). Domestic joint-venture (JV) refineries Satorp and Yasref added 400k bpd each of capacity in 2H14. With the completion of the fully-owned Jazan 400k bpd refinery in 2018/19, overall total refinery capacity would stand at 5.7mn bpd, of which Saudi Aramco's share would be 3.3mn bpd. This would allow Saudi Arabia to guarantee security of demand for its heavy crude, increase domestic job creation and diversification, build an export base of refined products (likely middle distillates to the EU) and lower imports of refined products.

National Transformation Plan: many promises, few details

The National Transformation Plan (NTP) released by the Saudi government is a further step towards execution of Saudi Vision 2030 through a set of ambitious interim targets to achieve by 2020 by 24 government bodies. The NTP is a comprehensive program encompassing medium-term growth boosting initiatives and fiscal consolidation measures. The sequencing and details of the fiscal measures are however still left nebulous. Implementation of growth-boosting initiatives will require a challenging crowding in of private sector investment, particularly from domestic sources.

The NTP is likely to introduce a number of one-off supply-side shocks to inflation through its fiscal or labor measures, which could negatively impact Real Effective Exchange Rate (REER) competitiveness. Discretionary consumption growth and real incomes are likely to remain under pressure. We find the diversification measures encouraging, although they appear too ambitious to be met within the time frame envisaged in the NTP, in our view. The focus on diversification could introduce medium-term incentives to adopt a more competitive exchange rate, while also detracting from the binding fiscal constraint under which the government has to operate. Energy policy is likely to be less aggressive and support oil prices. Fiscal targets are difficult to reach, but medium-term oil prices of at least US$50-65/bbl could support NTP progress.

Table 4: Selected macro Key Performance Indicators (KPIs) under the National Transformation Plan

Government entity	Key Performance Indicator (KPI)	Unit	Baseline	2020 target	Regional b'mark	Global b'mark
Ministry of Finance						
	Total non-oil revenues	SARbn	163.5	530	10.9	691.0
	Budgeted salaries and wages	SARbn	480	456	N/A	N/A
	Salaries and wages as a percentage of the budget	%	45	40	30	12
	Approved projects according to criteria and timeline (% of total)	%	0	40	30	78
	Credit rating	-	A1	Aa2	Aa2	Aaa
	Government debt as percentage of gross domestic product (%)	%	7.7	30	35	54
	Total recorded non-oil assets	SARtrn	3	5	N/A	N/A
Ministry of Economy & Planning						
	Total revenue resulting from privatization projects	Under Study	Under Study	Under Study	N/A	N/A
	Value of water and electricity subsidy decrease	SARbn	0	200	N/A	N/A
	Percentage decrease in non-oil subsidy	%	0	20	N/A	N/A
	Value of private sector contribution to GDP	SARbn	993.3	Under Study	122.6	14,133.8
Ministry of Energy, Industry and Mineral Resources						
	Value of exports of non-oil commodities	SARbn	185	330	Under Study	Under Study
	Value of the mining sector's contribution to GDP	SARbn	64	97	13	262
	Local content in expenditure of public and private sectors	%	36	50	Under study	57
	Petroleum production capacity	mn bpd	12.5	12.5	3.8	11
	Dry gas production capacity	bn cf pd	12	17.8	5.7	16
	Refining capacity	mn bpd	2.9	3.3	1.1	1.9
Ministry of Labor and Social Development						
	Unemployment rate for Saudis	%	11.6	9	Under Study	5.8
	Cost of employment of Saudis compared to expatriates	%	400	280	N/A	N/A
	Proportion of female labor force	%	23	28	Under Study	Under Study
Ministry of Hajj and Umrah						
	Number of formal Hajj pilgrims (domestic and foreign)	mn	1.5	2.5	N/A	N/A
	Number of Umrah Pilgrims from abroad	mn	6	15	N/A	N/A
	Number of Umrah pilgrims (domestic and GCC nationals)	mn	Under Study	Under Study	N/A	N/A
Commission for Tourism and National Heritage						
	% contribution of tourism sector to GDP	%	2.9	3.1	4.9	5.4
Ministry of Civil Service						
	% decrease in the payroll and benefits expenditure	%	Under Study	20	Under Study	22
	% of workers reduction in the civil service sector	%	Under Study	20	Under Study	18
Saudi Arabian General Investment Authority						
	Foreign Direct Investment (FDI)	SARbn	30	70	45	481
	Time needed to issue work visas for new expat employees	Day	30	10	10	3
	Time needed to issue new business permits	Day	19	1	8	0.5

Source: National Transformation Plan

Ambitious diversification agenda is not without risks

Diversification initiatives will require material participation from the private sector. Vested interests and the large bureaucracy could act as a dampener on timely reform implementation going forward. The large size of the NTP program may mean some of its outcomes are internally inconsistent, in our view. Cultural and entertainment changes could be opposed by conservative or clerical factions. Slippage could occur due to execution risk or reform fatigue, particularly due to the socio-economic impact of fiscal consolidation. Any material changes to subsidies or wages could be difficult to implement in the absence of a social safety net. While fiscal consolidation measures are unprecedented in scope, they appear too ambitious or difficult to reach and may leave a financing gap. As such, the measures may still fall short if oil prices do not stabilize, which suggests a need for a less aggressive energy policy going forward, in our view.

Government leaders continue to show focus on delivery

The US visit of Deputy Crown Prince Mohammed bin Salman and finalization of various Memoranda of Understanding (MoUs) with prominent US corporates helps provide a highly visible anchor for foreign investment and instil business confidence in regards to government focus. It also helps boost the profile of the Deputy Crown Prince both domestically and internationally. This is in line with our view that the government is likely to target rapid implementation of several 'low-hanging' reforms to spearhead the program and boost confidence. Energy sector liberalization could also serve as another highly visible anchor to the reform program, in our view.

Table 5: Announced US corporates plans during the US visit of the Saudi Deputy Crown Prince

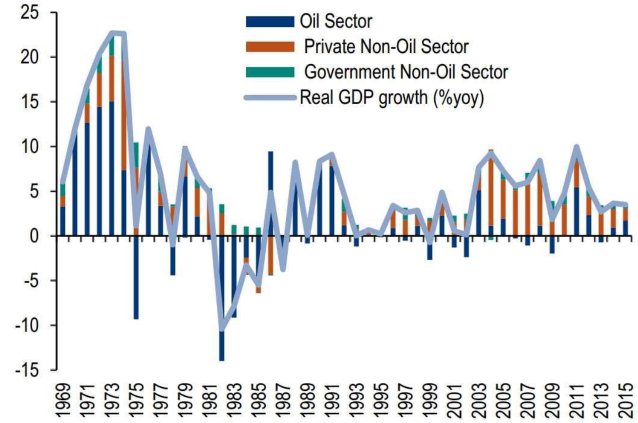
Company	Sector	Outcome
Six Flags	Entertainment	Company has been allowed to operate in Saudi Arabia
Pfizer	Pharmaceuticals	Company has been awarded a direct investment license
3M	Manufacturing	Company has been awarded a direct investment license
Dow Chemical	Chemicals	Company has been awarded a direct investment license
Microsoft	IT	Memorandum of Understanding signed
Cisco	IT	Memorandum of Understanding signed
Apple	IT	Discussions on entry to Saudi market possibly under way

Source: Press reports, Saudi Press Agency, BofA Merrill Lynch Global Research

Diversification drive is a step in the right direction

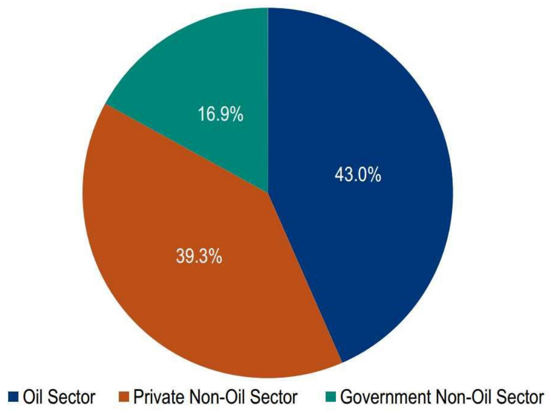
The comprehensive medium-term diversification drive introduced by the NTP follows the typical macro template, in our view. We discuss in a later section diversification lessons from Malaysia and Norway and conclude that the NTP contains elements from these successful case studies. In particular, it a) identifies key sectors with relatively high growth potential (mining, petrochemicals, manufacturing, retail and wholesale trade, religious and other forms of tourism, healthcare, real estate and finance); b), it starts to articulate supportive public sector industrial strategies, including through localization policies; and, c) it seeks to foster higher-value added in the economy through enhancements to processes, products and organizations.

Chart 5: More sustained private non-oil sector growth needed in Saudi



Source: Haver, BofA Merrill Lynch Global Research. Series uses contributions to growth from real GDP data with 1999 base year prior to 2010.

Chart 6: Oil sector large; government controls 60% of real economy

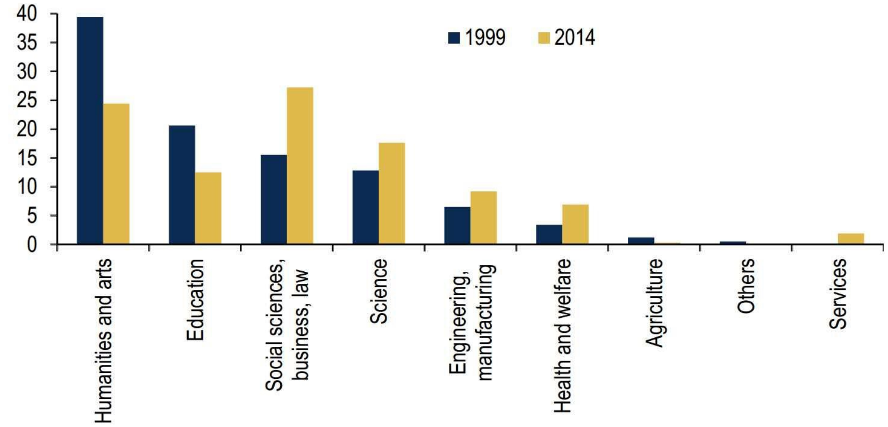


Source: Haver, BofA Merrill Lynch Global Research. Data as of 2015.

Human capital is critical and takes time to build

In our view, improvements to human capital, rather than business climate supply-side reforms, are the critical barriers to overcome. While building the appropriately qualified workforce takes time, Saudi Arabia has been making progress in this area. c25% of Saudi tertiary graduates complete humanities and arts programs, but this is down from 39.4% in 1999. Graduates from social sciences, business and law represented the largest share of graduates at 27.2% of total in 2014, up from 15.5% in 1999. Still, as we discuss in later sections, diversification prospects are mixed across sectors and will likely depend in part on providing the appropriate incentives to the private sector.

Chart 7: Distribution of Saudi tertiary graduates by field of study (%)

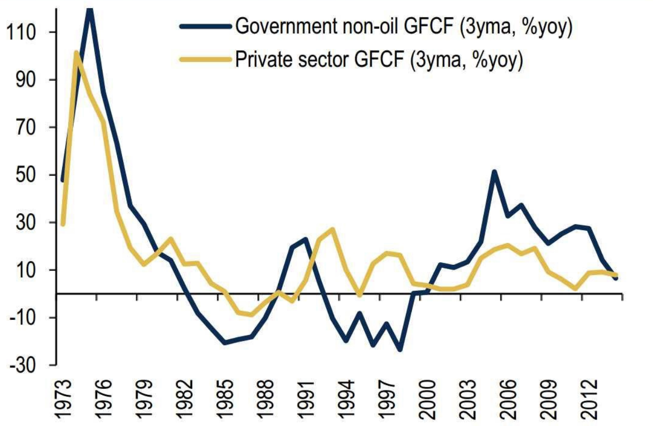


Source: UNESCO, BofA Merrill Lynch Global Research

Mobilizing domestic resources is key

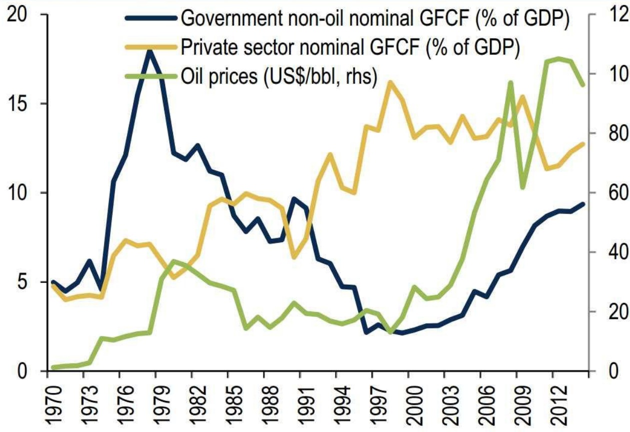
The NTP makes clear that part of the NTP costs (40%) will need to be shouldered by the private sector. Out of a total estimated NTP cost of SAR447bn (US$119bn; 18.4% of GDP), the NTP implies that the private sector will need to contribute about SAR179bn over the coming five years (US$47.6bn; 7.4% of GDP). Implementation of growth-boosting initiatives will require a challenging crowding in of private sector investment. The NTP expects only a SAR40bn increase in the level of Foreign Direct Investment (FDI), meaning that the bulk of private sector investment could have to come from domestic sources. While the government investment is generally in retrenchment mode, the NTP may be suggesting that selective and strategic projects are likely to go ahead. Authorities have also suggested a number of projects could be structured as Public-Private Partnerships (PPP).

Chart 8: Saudi government non-hydrocarbon and private sector investment growth are relatively correlated



Source: Haver, BofA Merrill Lynch Global Research. GFCF refers to Gross Fixed Capital Formation.

Chart 9: Crowding in domestic private sector investment when the government sector retrenches can be challenging

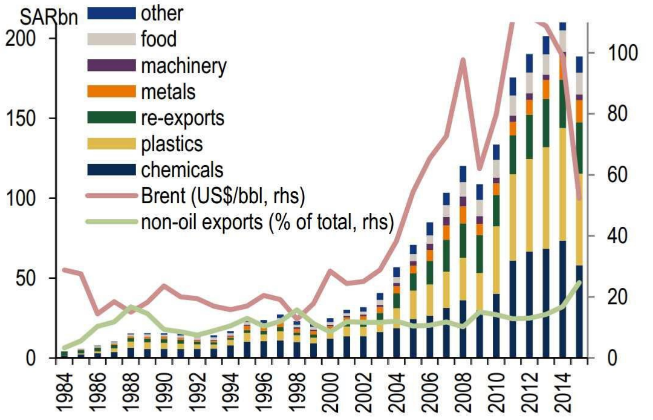


Source: Haver, BofA Merrill Lynch Global Research. GFCF refers to Gross Fixed Capital Formation.

Ambitious non-oil export and tourism targets, but little details

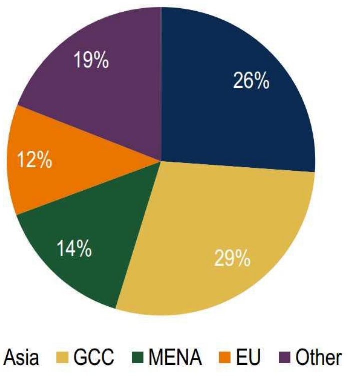
The NTP targets increasing non-oil exports by SAR145bn (US$38bn, 6.0% of GDP) to SAR330bn (US$88bn, 11.4% of 2020f GDP) over the next five years. Although implementation details are lacking, we anticipate that part of the increase could be linked to the higher mining output targeted. These would nevertheless be dependent on the global economic cycle. Higher exports of refined oil products, if all of the additional capacity to be installed by 2020 is exported, could add cUS$7.3bn to exports at current price levels (but not qualify towards non-oil exports targets under the NTP as they are classified as hydrocarbon exports). Currently, 61% of Saudi non-oil exports represents chemicals and plastics, and maintains a correlation with oil prices, and re-exports account for another 17% of total non-oil exports and have little added-value.

Chart 10: Chemicals, plastics, re-exports form bulk of non-oil exports



Source: Haver, BofA Merrill Lynch Global Research.

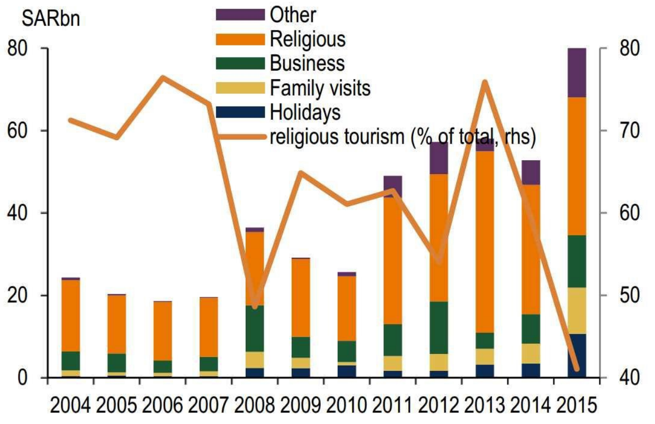
Chart 11: Non-oil exports are diversified in terms of destination



Source: Haver, BofA Merrill Lynch Global Research.

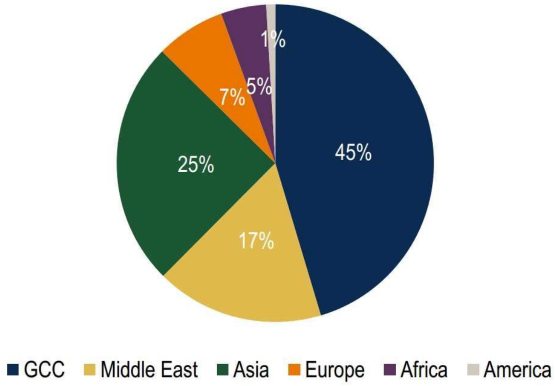
Likewise, the focus on religious tourism is appropriate given its importance in Saudi Arabia but NTP targets appear ambitious to us with a targeted 20%yoy CAGR increase in Umrah pilgrims. Religious tourism accounted for c40% of total tourist expenditure in Saudi Arabia in 2015, bringing in proceeds of SAR33.4bn (US$8.9bn; 1.4% of GDP). Note that, for balance of payment purposes, total tourism revenues stood instead at SAR37.9bn (US$10.1bn; 1.6% of GDP). The NTP targets would thus imply religious tourism external revenues/expenditures to increase to US$10-US$20bn (1.3-2.6% of 2020f GDP). This would be helpful on the external front but not a game changer on its own, in our view.

Chart 12: Breakdown of expenditure on inbound tourist trips



Source: SAMA, BofA Merrill Lynch Global Research.

Chart 13: Inbound tourist trips to Saudi Arabia by country of origin



Source: SAMA, BofA Merrill Lynch Global Research. Data as of 2015.

Mixed prospects for fiscal consolidation

The apparent redistribution of part of the proceeds of the flagship fiscal measures towards funding the NTP initiatives suggests that the fiscal consolidation efforts could fall short of the necessary requirement to narrow imbalances materially in the absence of a sustained oil price recovery. We calculate that the NTP targets a total net cumulative fiscal consolidation effort of SAR648bn (US$172.8bn) or c1% of GDP annually, assuming progressive and full implementation of the non-oil revenue targets.

Energy policy needs to support economic transformation

Energy policy is likely to be less aggressive going forward, particularly as the NTP suggests a constant oil production capacity as well as an increase in gas production domestically which could free up domestic crude for exports. To remain consistent with the targeted government debt accumulation path, we estimate that oil prices have to average at least US$50/bbl in 2016-20 along with no growth in spending from 2016 levels (excluding the additional cost of NTP initiatives but including implementation of the NTP non-oil revenue measures). Oil prices of at least cUS$65/bbl would be required if spending is not disciplined or revenue targets are missed by c50%, in our view.

Blueprint for fiscal measures lacks details

The approved NTP initiatives will have a total fiscal cost of SAR268bn (US$71.5bn; 11.6% of GDP) spread over five years (annually, SAR54bn equivalent to US$14.3bn or 2.3% of GDP). Government financing could represent 60% of total funding needs for the NTP initiatives, with total NTP costs of SAR447bn (US$119bn; 18.4% of GDP). Against that, the NTP envisages raising SAR366bn in non-oil revenues and cut the wage bill by 5% or SAR24bn (US$6.4bn or 1% of GDP). Non-oil revenue is seen rising to SAR530bn (US$141bn, 23% of 2016f GDP) by 2020 from SAR163.5bn (7% of GDP).

Wage bill measure unclear but may carry profound social implications

In an unprecedented austerity measure, the share of the wage bill in total spending would fall by 5ppt to 40% (SAR456bn, from SAR480bn) by 2020. It is yet unclear how this could be achieved in practice. Furthermore, note that the MoF budget announcement estimated the wage bill in 2015 to stand at SAR450bn, blurring the target. Privatizations of government entities are likely to lead to a natural drop in the public sector number of employees and wage bill. Yet we note that the Ministry of Civil Service has two Key Performance Indicators (KPIs) directly linked to a decrease in the payroll and benefits expenditure, and to a reduction in workers in the civil service sector. This would likely be a contentious and unpopular measure to implement, and its implementation and success appear challenging to us.

Spending path is unclear and is flattish (excluding NTP costs) in the best case

The spending path targeted or assumed by the NTP is unclear and does not appear to match realized outturns. It may suggest that spending is likely to remain flattish, excluding NTP costs. However, the indicative level spending would settle at is around peak 2014 expenditures levels and well above the 2015 realized spending and 2016 budgeted levels. While 2015 spending could have been understated and may be revised higher, at least SAR88bn from the 2015 spending is non-recurrent as it relates to the one-off Saudi Royal edicts decreed upon King Salman's accession.

Going by the indicated 2020 target for the share of the wage bill in total spending, this would suggest that government spending is targeted at SAR1,140bn in 2020. This represents a 2.7% increase over peak 2014 spending levels and 16.5% increase over 2015 actual spending levels. Excluding the NTP annual costs, this suggests that government spending should still increase by 11% over 2015 levels and drop by 2% over 2014 spending in nominal terms.

In our view, given the lack of precise spending targets, one way of remaining consistent is to use the two NTP figures for wages and infer the NTP fiscal spending from the stated share of wages in total expenditures. This would suggest that fiscal expenditures would reach SAR1,140bn in 2020, from SAR1,067bn in the base year (unstated in the NTP). This is a 6.8% increase, but excluding the NTP costs, is just 1.8% higher. As such, in the best case scenario, we think the NTP suggests flattish spending, excluding NTP costs. If we instead take the NTP level of spending inferred to be the correct amount, this would materially widen the fiscal deficit by SAR190bn (8.0% of GDP) and would be deeply negative for the sustainability of the fiscal and Fx stance, in our view.

Fiscal adjustment burden shifts to raising revenue, oil prices and exports

The possible flattish spending implication of the NTP would tend to indicate that the bulk of the spending cuts are now beyond us. Government projects are likely to continue to be prioritized, likely according to their economic benefits and relevance to the NTP. Flattish spending will thus shift the burden of forthcoming fiscal adjustment on raising non-oil revenues, raising hydrocarbon exports and a sustained recovery in oil prices.

Table 6: National Transformation Plan-related spending across government entities (2016-2020)

	Cost (SARbn)	Cost (US$bn)	% of total
Ministry of Housing	59.2	15.8	22.0
Royal Commission for Jubail and Yanbu	41.6	11.1	15.5
Ministry of Education	24.4	6.5	9.1
Ministry of Health	23.1	6.1	8.6
Ministry of communications and IT	14.9	4.0	5.6
Ministry of Agriculture	13.9	3.7	5.2
Ministry of Water and Electricity	12.9	3.4	4.8
Saudi Commission for Tourism & National Heritage	10.5	2.8	3.9
King Abdulaziz City for Science and Technology	8.3	2.2	3.1
Ministry of Labor	7.9	2.1	3.0
General Presidency of Youth Welfare	7.8	2.1	2.9
Ministry of Transport	5.6	1.5	2.1
Ministry of Social Affairs	5.4	1.4	2.0
King Abdullah City for Atomic and Renewable Energy	5.2	1.4	1.9
Ministry of Commerce and Industry	4.3	1.1	1.6
Ministry of Municipality and Rural Affairs	4.2	1.1	1.6
Ministry of Finance	3.4	0.9	1.3
Ministry of Culture and Information	3.3	0.9	1.2
Ministry of Economy and Planning	3.3	0.9	1.2
Ministry of Justice	3.2	0.9	1.2
Ministry of Petroleum and Mineral Resources	2.7	0.7	1.0
Saudi Arabian General Investment Authority	1.1	0.3	0.4
Others	2.3	0.6	0.9
Total	268.4	71.6	100.0

Source: National Transformation Plan

Focus on increasing non-hydrocarbon fiscal revenue could fall short of target

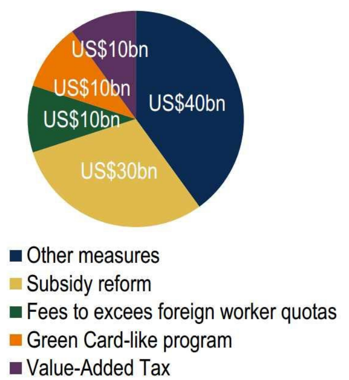
The breakdown of measures to raise non-oil revenues is not given. However, the tally is close to the US$100bn targeted in the Vision 2030 and is thus likely to include measures discussed then: hikes to administered energy prices, review of current levels of fees and fines, introduction of new fees, land tax (we estimate annual revenues of 1.5-2% of GDP in its first phase when fully implemented). In this regard, the NTP suggests forthcoming privatizations, imposition of taxes on harmful products (likely tobacco and soft drinks), introduction of a Value-Added Tax (VAT) for which we estimate annual revenues of up to 2% of GDP, implementation of a unified income tax and of an income tax on residents (although official comments later clarified there were no plans to tax nationals but stayed equivocal regarding expatriates).

Although there is no direct budgetary impact, water and electricity subsidies are targeted to decrease by SAR200bn. Non-oil subsidies are targeted to decrease by 20% by 2020. The latter, we believe, consists largely of subsidies for social and sports clubs, private education, private hospitals, and other agricultural subsidies. They were budgeted at SAR3.9bn (US$1.0bn; 0.2% of GDP) in 2016, down from SAR15.0bn (US$4.0bn; 0.6% of GDP) in 2015. A 20% cut to these on-budget non-oil subsidies would thus generate minimal savings of US$0.2-US$0.8bn (0.1-0.5% of GDP) depending if the reference base year was 2015 or 2016.

According to the Deputy Crown Prince pronouncements early in the year, medium-term plans are also likely to target the introduction of US$400bn of unutilized state assets (land, etc) to state-owned funds, with the latter in turn in charge of developing them into projects and companies than can be IPOed to the public. In our view, this may be linked to the Ministry of Finance assigned KPI of increasing total recorded non-oil assets (real estate, etc) from SAR3trn to SAR5trn.

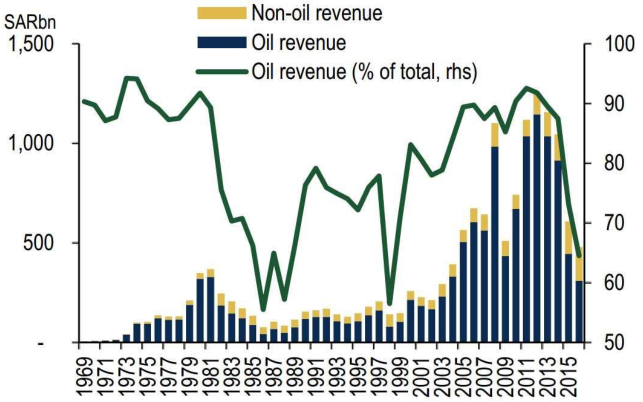
The Deputy Crown Prince provided a breakdown of the government target of raising US$100bn in additional non-oil revenue annually by 2020. The bulk of this increase in non-oil revenue would come from the restructuring of subsidies which could generate US$30bn per year (4.9% of GDP), according to the Deputy Crown Prince. VAT implementation, a Green Card-like program and a plan to allow corporates to hire foreign workers in excess of their official quotas could bring in US$10bn (1.6% of GDP) a year by 2020, according to the Deputy Crown Prince.

Chart 14: US$100bn in additional non-oil revenue targeted by 2020



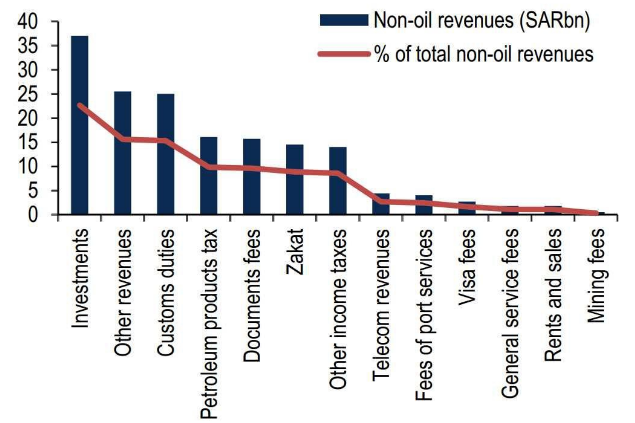
Source: Bloomberg

Chart 15: Government to focus on raising non-hydrocarbon revenue



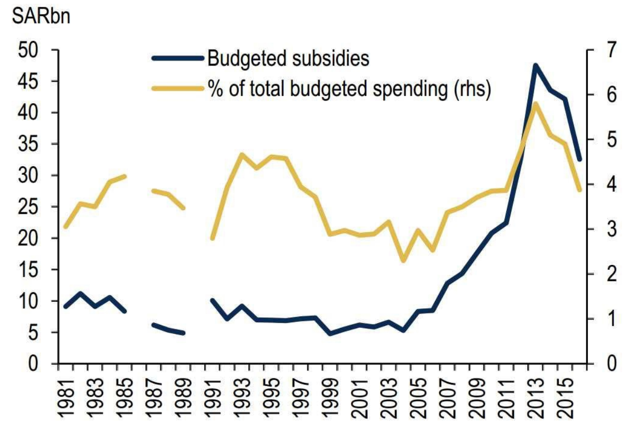
Source: Haver, BofA Merrill Lynch Global Research

Chart 16: Breakdown of non-hydrocarbon fiscal revenues



Source: MoF, BofA Merrill Lynch Global Research. Data as of 2015.

Chart 17: Budgeted non-oil subsidies form a small portion of spending



Source: SAMA, MoF, Bank of America Merrill Lynch Global Research.

Further energy subsidy reform is likely

The figures announced by the Deputy Crown Prince regarding proceeds of subsidy reform suggests material changes over the next 3-4 years, with annual savings close to the landmark first round of administered price changes in late 2015. However, his suggestion that the changes could be accompanied by partially offsetting cash transfers to poor households suggests policy-making caution. We do not think such a redistributive system could be technically put in place in a short time span.

Recall that concurrently with the 2016 budget, a first round of energy subsidy reform saw sweeping energy, water and electricity administered price changes being instituted in late December. We estimated the natural gas price hike on petrochemical firms, domestic crude oil price hike as well as the combined gasoline and diesel price hike introduced could add US$2.2bn, US$2.0bn and US$3.8bn to central government revenues if fully passed to the budget (a combined 1.2% of GDP). Impact of continued similar policies on domestic prices would be a further 1.5-2ppt annual increase in headline inflation, and a gradual squeeze to household incomes (where gasoline, water and electricity likely represented c.1.5%, c0.4% and 1.6% respectively of consumer spending basket prior to the late December 2015 administered price adjustments).

Table 7: Administered energy price adjustments carried out in December 2015

Product	Price prior to Dec 2015	Current price	% change
Methane (US$/mn BTU)	0.75	1.25	66.7
Ethane (US$/mn BTU)	0.75	1.75	133.3
Arab light (US$/bbl)	4.24	6.35	49.8
Arab Heavy (US$/bbl)	2.67	4.4	64.8
Diesel (US$/bbl)	9.8	14 - 19.1	68.5
Heavy Fuel Oil (HFO) 380cst	2.08	3.8	82.7
Heavy Fuel Oil (HFO) 180cst	2.08	4.25	104.3
Propane (USD/metric ton)	-	-	-
Naphtha (USD/metric ton)	-	-	-
Butane (USD/metric ton)	-	-	-
Natural gasoline (USD/metric ton)	-	-	-
Kerosene (US$/bbl)		25.7	
95 octane gasoline (SAR/ltr)	0.6	0.9	50.0
91 octane gasoline (SAR/ltr)	0.45	0.75	66.7

Source: SPA, BofA Merrill Lynch Global Research. Propane, naphtha, butane and natural gasoline used to be: price using the following formula: Japanese Naphtha price less transportation from Saudi multiplied by 0.72. The formula now is differentiated by product and is as follows: Japanese price of the underlying product less transportation from Saudi multiplied by 0.8.

Fiscal consolidation plan looks challenging to fully achieve

The US$100bn target in additional non-oil revenue through 2020 appears difficult to reach and is likely to leave a sizeable financing gap based on the current proposals, in our view. This would thus imply that fiscal consolidation is likely to fall short of narrowing imbalances materially without a sustained oil price recovery.

The fees on excess foreign worker quotas and the green-card like program targeted proceeds are difficult to reach, in our view. We calculate that to raise US$10bn from a green-card like program, authorities will need to set the program fees at US$1,500 (in line with upper bound of similar programs elsewhere) and have all expatriate workers applying to enrol in the program. This appears an unrealistic target to meet, particularly as it is not clear how these proceeds would be recurrent yearly. As for the foreign worker quotas proceeds, we calculate that, based on private sector expatriate figures and the stated 85% overall compliance with Nitaqat sector quotas, that the fees would need to be set at a large cUS$10,500 per non-compliant worker to be consistent with the NTP target. The large fee amount would likely be a material disincentive to private sector corporates and a squeeze on their profits.

One of the obstacles to meeting the subsidy reform target is likely to be political in nature given the impact on inflation and household incomes. We calculate nevertheless that the target could be reached through a combination of a administered prices changes. We do not think there is major room to increase natural gas feedstock prices to petrochemical firms, given the NTP focus on the petrochemical sector as a source of future growth. We estimate raising natural gas prices to US spot levels in an oil price environment of US$50/bbl would bring in revenues to the central government of just US$2.5bn (0.4% of GDP) if fully passed to the budget. This is likely to suggest that the bulk of the future adjustments are likely to center on hikes to administered price adjustments for domestic crude oil sales, diesel and gasoline. Selling domestic crude at US$25/bbl rather than an average of US$5.4/bbl would bring in US$20.8bn in revenues. Gasoline and diesel prices could rise by the same percentage as in December 2015 to bring in additional revenues for a minimum inflation pick-up cost of 0.5ppt.

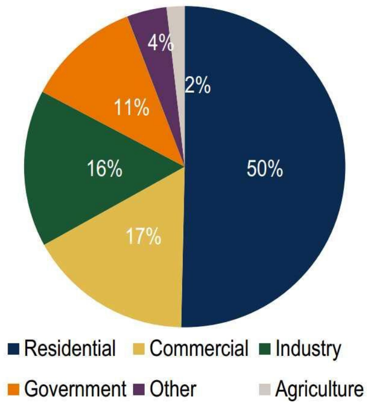
The most difficult part would be meeting the US$40bn in other non-oil revenues target, in our view. Given that the latter's breakdown was not provided, we have attempted to group several measures being studied (according to local press) to provide a tentative decomposition. Still, our analysis suggests the need for a further US$16.8-US$20.7bn in financing to meet the US$40bn target. Land tax proceeds of US$11bn in the first phase of implementation are the largest item, in our view. We calculate that a 20% sin tax on tobacco and sugary drinks would raise US$2.1bn (0.3% of GDP) at a 0.5ppt cost to inflation. A remittance tax being studied, according to local press, and could impose a tax of 6% gradually decline to 2% over a number of years. At the upper bound of the tax rate (6%), we calculate that the remittance tax would raise US$2.3bn (0.4% of GDP). We calculate that a 10% income tax on expatriates would raise at least between US$3.9-US$7.8bn (0.6-1.2% of GDP), depending on the assumed annual overall expatriate earnings. That being said, measures to tax expatriates could be detrimental to diversification and labor force prospects and were opposed to politically in the past.

Table 8: Fiscal consolidation measures through 2020 leave undisclosed financing gap

Fiscal measure	Annual revenue raised (US$bn)	Annual revenue raised (% of GDP)	Impact on inflation (ppt)	Remark
Non-oil revenue measures	100	15.8	-	NTP target
Fees on excess foreign worker quotas	10	1.6	-	NTP target
Green-card like program	10	1.6	-	NTP target
VAT	10	1.6	5	NTP target
Subsidy reform	30	4.7		NTP target
Natural gas feedstock for petrochemicals	2.5	0.4	-	BofA ML assumption
Domestic crude	20.8	3.3	-	BofA ML assumption
Gasoline	1.4	0.2	0.5	BofA ML assumption
Diesel	1.9	0.3	-	BofA ML assumption
Other measures	40	6.3	-	NTP target
Land tax	11	1.7	-	BofA ML assumption
Remittance tax	2.3	0.4	-	BofA ML assumption
Income tax on expatriates	3.9-7.8	0.6-1.2	-	BofA ML assumption
Sin tax on tobacco	0.7	0.1	0.1	BofA ML assumption
Sin tax on sugary drinks	1.4	0.2	0.4	BofA ML assumption
Memo:				
Undisclosed revenue financing gap	16.8-20.7	2.7-3.3	-	BofA ML assumption
Cost savings	**Annual savings (US$bn)**	**Annual savings (% of GDP)**		
20% cut to on-budget non-oil subsidies	0.2-0.8	0.1-0.5	-	BofA ML assumption
5% cut to the wage bill	6.4	1.0	-	NTP target
Cost increases	**Annual cost increase (US$bn)**	**Annual cost increase (% of GDP)**		
NTP costs (US$71.5bn)	14.3	2.2		NTP target

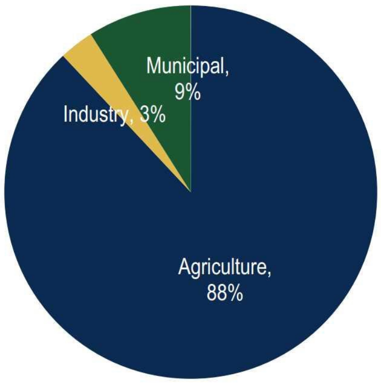
Source: National Transformation Plan, BofA Merrill Lynch Global Research

Chart 18: Breakdown of electricity consumption by sector



Source: SAMA, BofA Merrill Lynch Global Research. Data as of 2015.

Chart 19: Water consumption by sector

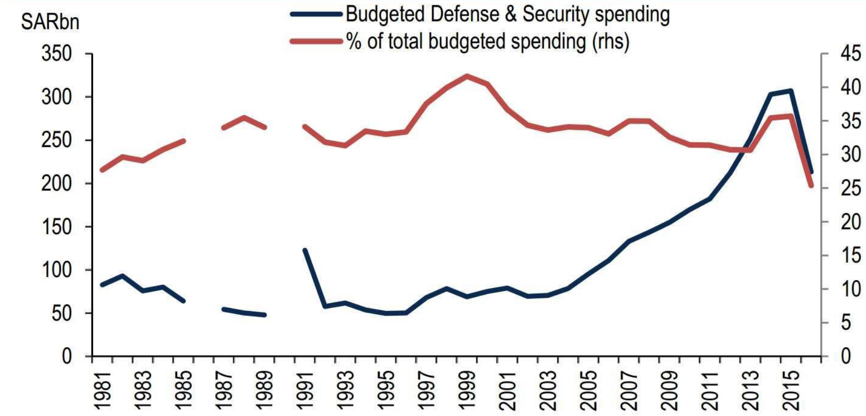


Source: UN, BofA Merrill Lynch Global Research. Data as of 2010.

Budgeted defence spending – up or down

Localization policies in the defence industry could help cut imports (total defence imports of US$9.8bn equivalent to US$1.5bn or 6% of total imports in 2015, according to consultancy IHS) and conserve Fx reserves, but we expect this to be a slow and gradual process. We think that large defense imports are typically amortized over a number of years. Defence spending is an important part of fiscal spending in Saudi Arabia, accounting for c30% of total budgeted spending. Defence spending in 2016 was budgeted at SAR213bn (US$56.9bn; 9.0% of GDP), down from SAR307bn (US$81.8bn; 12.7% of GDP) in 2015. Increased military and security projects in 2015 saw an additional overspend of SAR20bn last year. It is unclear to us if all of the regional military costs have been recognised on-balance sheet. A negotiated settlement to the Yemen conflict through the ongoing political negotiations in Kuwait could help contain security spending near-term.

Chart 20: Budgeted spending – up or down?



Source: SAMA, Ministry of Finance, Bank of America Merrill Lynch Global Research.

Privatization comes back to the government's agenda

The consideration of a privatization program is not surprising at the current juncture, in our view. Recall that the Saudi privatization program was initially started in 1999 with the creation of the now-dissolved Supreme Economic Council, following the drop in oil prices in 1998. Small-scale privatizations took place in the early 2000s, and 2002 saw large privatization in the telecommunication sector and postal services.

Selling public sector stakes is one of the non-debt creating financing options for the government that would help minimize the direct drain on Fx reserves, encourage private sector development and improve services delivery. That being said, given the lack of non-oil taxation, this would only contribute to one-off financing flows alongside savings from a drop in budgetary allocations to the privatized entities, in our view.

The 2015 budgetary appropriations for public institutions totalled SAR163.7bn (US$43.7bn; 19.0% of total budgeted spending and 6.8% of GDP). We calculate that the public institutions related to the entities that appear to have been slated for privatization according to press reports account for budgetary appropriations of SAR124.8bn (US$33.3bn; 5.2% of GDP. This is likely to be the upper bound of fiscal savings possible under the privatization program. In our view, this is unlikely to be realized fully as it includes a large number of universities and because most timelines appear to center around 2020.

Table 9: Privatization / Public Private Partnerships (PPPs) planned

Entity	Date	Comment (press reports and Saudi Vision 2030/NTP, unless stated otherwise)
Airports	2016-2020	Targeted sale of 11 airport units by 2020
Saudi General Grains Organisation	1Q17	A financial advisor has been appointed
Saudi Aramco	by 2018	A stake of less than 5% could be sold
Stock exchange (Tadawul)	by 2018	A financial advisor has been appointed
Saudia Medical Services	-	Saudi Arabian Airlines appointed a financial adviser for the privatization of its unit Saudia Medical Services and had its Board of Directors restructured in June
Saudi Electricity (SEC)	by 2020	SEC's generation assets are likely to be split into four separate regional companies where minority stakes would be sold to major global utilities or sold in the public market
Saline Water Conversion Corporation (SWCC)	-	Investment partners are likely to be sought to buy a stake in production assets, with the holding company to be listed later
General Port Authority	by 2020	Commercialization program to be completed by 2020
Saudi Post	by 2020	Saudi Post to be turned into a holding company with six subsidiaries. The NTP aims to transform it to a commercially viable company with government subsidies (SAR2bn) to be phased out by 2020.
Education / schools	-	Official pronouncements have suggested this could be considered, but there does not appear to be concrete plans for now
Healthcare / hospitals	-	The 2030 vision suggests no privatisations in the near-term, in our view. Authorities are likely to look to improve management and quality of service before considering privatisation, in our view.
Road, railway and port projects	-	NTP suggests PPPs are planned with percentage in private sector contribution to development and operation to increase

Source: Press reports, Saudi Vision 2030, National Transformation Plan

Table 10: Public institutions budget appropriations

	SARbn	US$bn	% of GDP
Universities	55.7	14.8	2.3
Saudi Arabian Airlines	28.5	7.6	1.2
Saline Water Conversion Corporation	15.6	4.2	0.6
General Authority of Civil Aviation (GACA)	15.5	4.1	0.6
Saudi Post Organization	3.2	0.8	0.1
Grain Silos and Flour Mills Organization	2.9	0.8	0.1
Saudi Ports Authority	1.8	0.5	0.1
Saudi Railways Organization	1.7	0.4	0.1
Other	38.9	10.4	1.6
Total	163.7	43.7	6.8
of which, public institutions in sectors that could be privatized	*124.8*	*33.3*	*5.2*

Source: SAMA, Ministry of Finance, Bank of America Merrill Lynch Global Research. Data as of 2015 budget.

Sales of PIF assets could help replenish fiscal reserves but not Fx reserves

Secondary sales of domestic assets could be a faster way than Initial Public Offerings (IPOs) and privatizations for the Ministry of Finance (MoF) to raise its fiscal reserves at SAMA, in our view. We look at this possibility to assess the potential for the MoF to use it as an option to support its fiscal reserves at SAMA and prevent a debt build-up in response to ongoing fiscal deficits.

If this were to take place, this could occur through holdings of the PIF rather than the pension funds who are large institutional investors in the domestic equity market, in our view. We calculate that the PIF could sell stakes worth around SAR200bn (US$53.6bn) domestically through liquidating existing minority stakes and selling down majority holdings while retaining control. If the PIF holdings are above 50%, we assume arbitrarily that the holding is strategic, and thus calculate the current value of the stake that the PIF (using the current stock price in the market) can sell down so it retains a controlling 51% ownership share. If the PIF holdings are below 50%, we arbitrarily assume that the holding is not strategic, and thus calculate the current value of the stake that the PIF holds (using the current stock price in the market) as we assume it could be fully liquidated.

The proceeds of sales domestically (without foreign participation) would increase central government deposits at SAMA, decrease other domestic liabilities of SAMA but keep Fx reserves unchanged. Because of the latter, because such large coordinated sales could weigh on the market and as the PIF is targeted to become the largest SWF by Saudi authorities, we doubt this is going to be a course policy-makers are likely to be considering in the near term.

Table 11: Public Investment Fund listed domestic assets and their monetization potential in regards to replenishing fiscal reserves at SAMA

Entity	Sector	Current government stake (%)	Free-float (%)	Potential value of secondary offering if a government stake is sold, but with the government still retaining 51% control (US$bn)	Potential value of secondary offering if government's existing minority stake is liquidated (US$bn)
Saudi Basic Industries Corporation (SABIC)	Chemicals	PIF: 70.0%; GOSI: 5.7%	24.3%	12.3	0.0
Saudi Telecom (STC)	Telecom	PIF: 70.0%; GOSI: 7.0%; PPA: 6.77%	16.2%	6.5	0.0
Saudi Electricity (SEC)	Power	Government: 74.3%; Aramco: 6.9%	18.8%	5.1	0.0
Saudi Ground Services	Travel	Saudi Arabian Airline: 52.5%; National Aviation: 14.7%	30.0%	0.4	0.0
Saudi Real Estate Company	Real estate	PIF: 64.57%	35.5%	0.1	0.0
National Commercial Bank (NCB)	Banks	PIF: 44.3%; GOSI: 10%; PPA: 10.04%	35.7%	0.0	9.5
Maaden	Mining	PIF: 49.99%; GOSI: 7.98%; PPA: 7.45%	34.6%	0.0	5.9
Samba	Banks	PIF: 22.91%; GOSI: 11.76%; PPA: 15.04%	73.2%	0.0	2.4
Yanbu National Petrochemical Co (Yansab)	Chemicals	SABIC: 51%; GOSI: 11.92%	37.1%	0.0	2.2
Saudi Arabian Fertilizer Company (SAFCO)	Chemicals	SABIC: 42.99%; GOSI: 12.2%	44.8%	0.0	2.0
Riyadh Bank	Banks	PIF: 21.75%; GOSI: 16.72%; PPA: 9.18%	46.6%	0.0	1.9
Southern Province Cement	Cement	PIF: 37.43%; GOSI: 15.82%	46.8%	0.0	1.1
National shipping company	Transport	PIF: 22.55%	80.0%	0.0	0.9
Saudi Catering	Transport	General Airline Services KSA: 35.7%	39.8%	0.0	0.8
Kayan	Chemicals	SABIC: 35%	65.0%	0.0	0.6
Alinma Bank	Banks	PIF: 10%; GOSI: 5.10%; PPA: 10.71%	74.1%	0.0	0.5
Qassim cement	Cement	PIF: 23.35%; GOSI: 15.09%; PPA: 5.67%	79.2%	0.0	0.4
Yanbu cement	Cement	PIF: 10%; GOSI: 12.37%	68.5%	0.0	0.2
National Agriculture Development	Agriculture	PIF: 20%	85.1%	0.0	0.1
Saudi Fisheries	Agriculture	PIF: 39.99%	38.5%	0.0	0.1
Saudi Public Transport	Transport	PIF: 15.72%	100.0%	0.0	0.1
Eastern Province Cement	Cement	PIF: 10%; GOSI: 10.65%	89.4%	0.0	0.1
National Gas and Industrialization	Chemicals	PIF: 10.91%	88.0%	0.0	0.1
Saudi Ceramic	Consumer	PIF: 5.94%; GOSI: 16.19%	83.8%	0.0	0.0
Petro Rabigh	Chemicals	Aramco: 37.5%; Sumitomo: 37.5%	25.0%	0.0	0.0
Total				24.3	28.6

Source: Bloomberg, Bank of America Merrill Lynch Global Research. PIF = Public Investment Fund. GOSI = General Organization for Social Insurance. PPA = Public Pension Agency. Based on stock prices as of 28 June 2016.

Material fiscal consolidation would be necessary to narrow macro imbalances

There is no substitute for fiscal consolidation to maintain Fx policy unchanged. We map out below paths for several macro variables depending on fiscal policy and oil prices, with higher oil prices easing the adjustment requirement.

In a worst case scenario where US$25/bbl oil prices would persist for the next five years, sustainability of the Saudi Fx policy rests on material fiscal consolidation, which we believe is achievable based on the mix of revenue-raising and spending restraint plans likely or announced to date. This would require a total 5-year cumulative adjustment of SAR500bn (US$133bn; 20% of 2015 GDP or 4ppt of GDP annually) from 2016 onwards, which is close to what the NTP appears to be targeting in non-oil revenue measures (US$100bn, excluding NTP costs). We think 75% of the adjustment can take place through revenue-raising measures. The remainder of the adjustment would need to take place through capex cuts and is similar in size to the capex retrenchment of the 1980-90s. However, note that if one assumes that the 2015 budget deficit will be revised higher to 18.9% of GDP, this adds a need for a further 4ppt of GDP cumulative fiscal adjustment to maintain the same trajectory of Fx reserves.

An overhaul of fiscal policy will likely be required

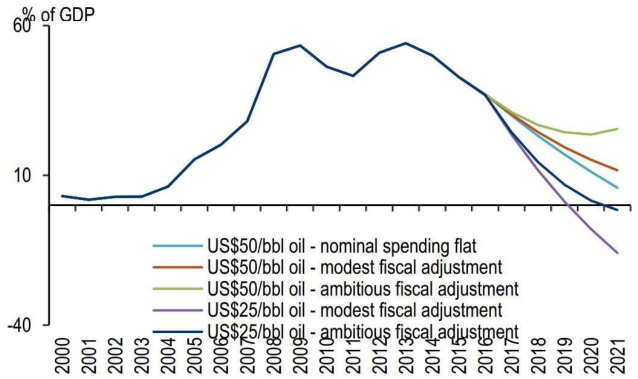
In line with the 2016 budget announcement, we would expect tight budgets to be passed. We expect a VAT with 5% yield to be implemented in 2018, alongside further administered price adjustments for energy, water and electricity. Implementation of the land tax is likely, but it is unclear whether the proceeds will be ring-fenced to be used solely for housing projects. Using solely the recurring existing proceeds of the December administered price adjustments and a VAT introduction, we still see a need for further fiscal restraint in the order of 1ppt of non-oil GDP annually. The latter would be equivalent to a SAR100bn cumulative spending cut, which we believe can be accommodated from capex retrenchment. On the basis of the above, fiscal consolidation would likely drive real GDP growth lower to an average of just around 1% of GDP. This estimate is based on a ST GCC fiscal multiplier of 0.2 for real spending to real non-oil GDP, in line with academic literature.

Fx policy stays the course; NTP blurs medium-term incentives

We do not believe Deputy Crown Prince Mohammed bin Salman is considering Fx reform in the near-term. We have expressed here our view that the Fx peg would hold as fiscal policy undergoes a sizeable multi-year adjustment. In a bearish scenario for oil prices, the NTP is unlikely to be implemented as deep budget cuts are likely to be required to sustain the Fx policy. We believe an oil price of at least US$50/bbl is broadly consistent with full and timely NTP implementation, but fiscal policy still needs to be disciplined. If fiscal non-oil revenue targets are missed, a higher oil price is likely to be needed.

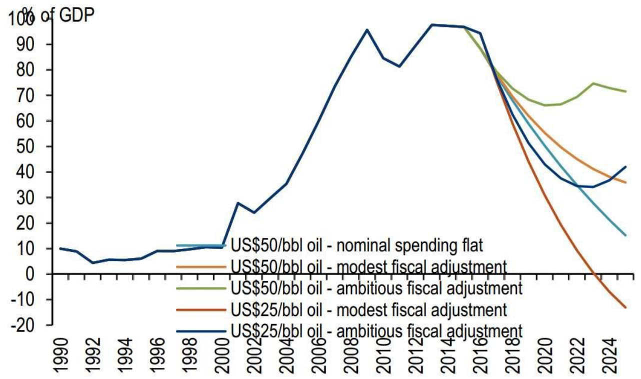
However, if the NTP is successful in diversifying the economy, it will make the costs of running an overvalued exchange rate more prohibitive, particularly as the external accounts are likely to become more responsive to an Fx shock. It also introduces a pro-growth bias for policy-makers in charge of diversifying the economy, which may conflict with the implementation of unpopular fiscal reforms or the depth of fiscal consolidation needed at weak levels of oil prices. The NTP diversification initiatives may suffer from a timing mismatch when it comes to net external receipts generated. This is because the sectors that would generate Fx receipts (non-oil exports, religious tourism) or localization policies are likely to be slow to ramp-up, while Fx demand in the economy is likely to be supported (rather than deflated) by government diversification policies. Last, the desire to build a large SWF may suggest a desire to conserve Fx reserves.

Chart 21: Illustrative paths of Saudi government deposits at SAMA



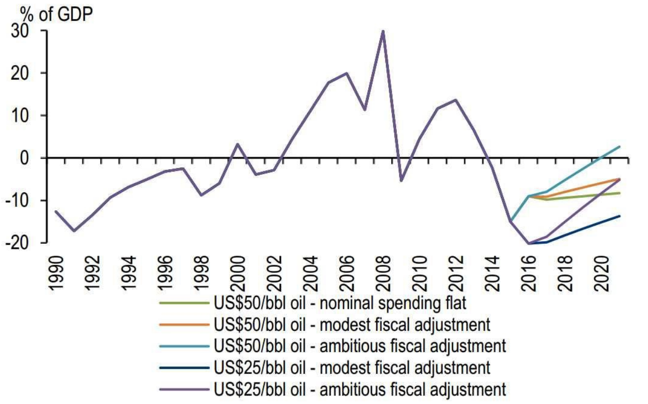
Source: , Haver, BofA Merrill Lynch Global Research estimates

Chart 22: Illustrative paths of SAMA foreign assets



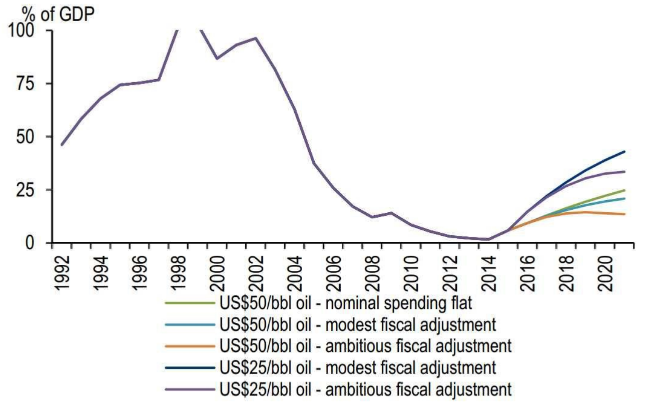
Source: Haver, BofA Merrill Lynch Global Research estimates

Chart 23: Illustrative paths of Saudi fiscal balance



Source: Haver, BofA Merrill Lynch Global Research estimates

Chart 24: Illustrative paths of Saudi government debt



Source: Haver, BofA Merrill Lynch Global Research estimates

Consumer to face mixed trends

Labour policies are likely to remain a major focus for the Saudi government. Low oil prices have dented the government's direct ability to support the consumer compared to the boom years. Furthermore, negative to flat public sector wage growth is likely to dampen income and consumption trends, although Specialized Credit Institutions (SCIs) could alleviate tightening banking sector household lending standards.

The NPT suggests a focus on boosting female employment as well as increasing the cost of foreign labor relative to national labor. The latter implies a negative impact on corporate margins and profits, in the event higher labor costs are not reflected in higher consumer prices and inflation). The negative near-term cost and efficiency implications of continued labour market reforms are mitigated by the fact that eventual increase in private sector employment of higher-paid Saudi labour should prove supportive for consumption trends once the dust settles.

The NTP targets increasing the cost of employment of Saudis compared to expatriates from 400 to 280. We calculate that the ratio of private sector wages of Saudis compared to non-Saudis stood at 3.64 and 1.29 for males and females respectively in 2015. The NTP target is thus akin to increasing non-Saudi private sector wages by 43% over the period to 2020, a CAGR increase of 7.4%. This would squeeze corporate profits, lead to inflationary pressures in the economy, and lead to annual additional Fx outflows through remittances of US$2.8bn (0.4% of GDP). As such, it is more likely that the employment cost of non-Saudis is likely to be increased through measures such as fees for government services, taxes and levies. It is unclear that the NTP target can be met solely through this route, but it would avoid additional Fx outflows despite still having a negative impact on corporate margins and inflation, in our view.

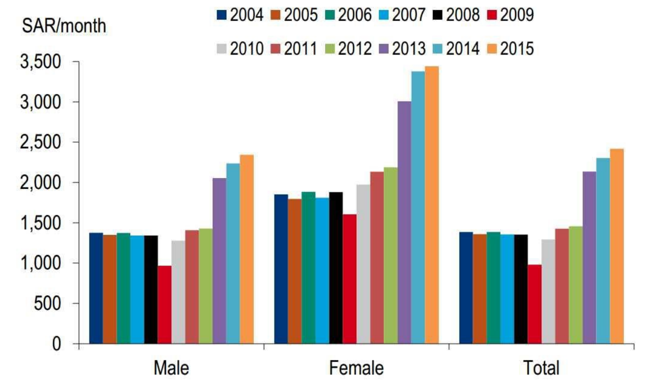
The Nitaqat Saudization program is likely to be extended in the near-future through the forthcoming launch of the "balanced Nitaqat" initiative. While employment quotas are likely to be tightened, they will also take into account wage levels, job quality and women employment ratios, according to local press. A likely target is likely to be increased job creation for Saudis in the retail sector through higher quotas, in our view.

Table 12: Selected drivers of Saudi consumer

	Direction	Comment
Oil prices	Volatile	Indirect impact through economic activity and confidence
Population growth	Decreasing	Direct impact through lower number of expatriates and national birth rate
Public sector wage growth	Flat to contracting	Freeze in government wage bill
Private sector wage growth	Flat to minor increase	Some sectors to contract (construction), others resilient (retail, staples)
Saudi private sector male employment	Mixed	Saudization helps, but economic slowdown and supply-side factors impact negatively
Saudi private sector female employment	Increasing	NTP aims to increase it through several measures
Consumer leverage	Decreasing	Banking sector liquidity and lending standards are tightening
Real estate finance	Increasing	NTP measures are supportive in this area
Specialized Credit Institutions (SCIs)	Flat	SCIs can continue to support the economy
Saudization	Increasing	Negative impact short-term, positive impact medium-term
Subsidies	Decreasing	Timing and breadth unclear as of yet
Other fiscal consolidation measures (VAT, etc)	Likely forthcoming	Direct and indirect negative impact on consumer

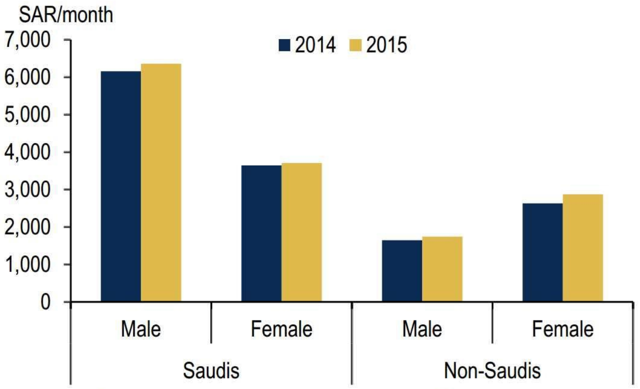
Source: BofA Merrill Lynch Global Research

Chart 25: Private sector average monthly wages are increasing



Source: SAMA, BofA ML Global Research. SAMA wage data coverage appears incomplete.

Chart 26: Wedge between Saudi and non-Saudi private sector salaries



Source: General Organization for Social Insurance (GOSI), BofA ML Global Research.

Banks and contractors will require supportive policies to realize NTP benefits

The banking sector is targeted to benefit from the NTP initiatives through greater disintermediation (SMEs, real estate financing, non-oil export financing) and greater household savings. However, domestic liquidity is likely to remain structurally tight, in our view. The NTP targeted increase in non-oil fiscal revenue is likely to tighten domestic liquidity, all else being equal, as it would have the effect of extracting resources from the non-hydrocarbon private sector (as opposed to government domestic spending that injects liquidity).

Linked to liquidity trends and banking sector asset quality, the issue of government arrears to contractors is likely to take centre stage, in our view. Resolving this is likely to be required if the capex projects under the NTP are to be executed on a timely fashion, in our view. We think the government's proposal to issue "I Owe You' notes (IOUs) could be a step in this direction.

Government looks to issue IOUs

According to the press, the Saudi government has paid contractors some cash on its arrears, and is considering to issue "I Owe You' notes (IOUs) to them. Contractors would receive these notes to cover their outstanding dues, which they could hold until maturity or sell on to banks. The government has been highlighting it wanted to settle the contractor sector dues, and recently a high-profile domestic contractor has made mass redundancies. In the previous downturn in the 1980s, arrears to domestic contractors were also likely to have been accumulated. We highlighted that the fiscal deficit was likely under-reported last year. Also, construction was one of the fastest growing sector last year both in terms of domestic activity and of domestic credit.

Government arrears to contractors could be significant

We estimate the range of unpaid capex for 2015 to be SAR23bn-SAR100bn (US$6bn-US$27bn or 1-4.3% of 2015 GDP). At the very minimum, unpaid capex could be around SAR23bn. This is the amount of additional loans taken out by construction contractors over 2015 compared to 2014. Total construction loans were SAR106bn in 2015 versus SAR83bn in 2014. Contractors likely did not go to obtain alternative bank funding for the same amount as the delayed payments from the government, suggesting delayed payments should be larger than SAR23bn, in our view.

By the same token, the 2015 budget outturns imply capex was down by 45% to SAR205bn versus 2014 levels of SAR370bn. This is of course unlikely as construction grew by 5% in real terms in 2015. Correcting the 2015 budget by the likely underreporting (an additional SAR100bn in spending, assuming it was all capex-related), then capex spending was likely down by SAR65bn (17%) in 2015. This would be in line with the c15%yoy drop in construction awards as reported by MEED.

In comparison, government overdues in the healthcare sector are much smaller. Based on the disclosures of two listed healthcare groups, we estimate the growth of government receivables owed to them alone is SAR0.4bn. We estimate the total sector government overdues to private healthcare operators could be 2-3 times this level.

IOUs to serve multiple purposes

IOUs could support simultaneously the multiple needs of the government, domestic banks and contractors, in our view. The government will be able to conserve fiscal reserves and restructure the maturity of its outstanding dues. Domestic banks would substitute corporate credit risk with sovereign risk and improve asset quality, while contractors could improve their liquidity and working capital position by cashing in early on the IOUs. We believe most of the contractors impacted could be domestic ones, and we presume these IOCs would be issued in domestic currency. Furthermore, to the extent these IOUs are structured as tradable instruments, this could better distribute risk in the financial sector to those agents more capable or willing to hold it.

Domestic liquidity could be eased, but Fx outflows may not subside

The impact of IOUs on domestic liquidity will likely depend on the issuance mechanism, of which we have little visibility for now, though this could likely ease domestic liquidity. We would however expect subsequent Fx reserves losses due to an increase in demand for Fx consequent to the liquidity injections (a 50% leakage would for instance lead to US$13.5bn loss in Fx reserves at the high-end of the estimated arrears range).

One possibility for the mechanism would be for banks to extend more secured credit to contractors using the IOUs as collateral. In this case, domestic liquidity is likely to remain tight as the banking sector balance sheet would stay stretched. Another, perhaps more likely, option would be for banks to accept IOUs as deposits. Contractors would then be able to transfer IOUs to their trade creditors and withdraw cash regularly, with the total amount being likely a portion of the face value of the IOUs. Banks would hold new government instruments on the asset side of their balance sheet. This mechanism could also be structured as a simple sale of the IOUs to domestic banks (with a haircut), which domestic banks would accommodate through equivalent changes on the asset side of their balance sheet, without the creation of additional liabilities.

In the latter option, to avoid straining the banking sector liquidity, the deposits withdrawal or in effect the increased money in circulation would likely require monetary accommodation from SAMA, in our view. We believe this could take the form of liquidity injection, possibly through repo operations for the IOUs. Although unconventional, this possible source of easing of domestic liquidity would still be consistent with Article 6 Saudi Arabia's Currency Law which imposes a 100% currency backing by Fx reserves, such that that currency issued cannot exceed foreign reserve assets. The full cover requirement would still be in force as reserve money currently represents just 14% of foreign reserve assets, down from c100% throughout the period 1996-2000.

Lessons in diversification

There are only few examples of countries that successfully managed diversification away from primary dependence. The examples of Malaysia and Norway suggest that a sound institutional framework, robust fiscal framework, supportive public sector industrial policies, appropriate business climate and human capital are important factors in enabling successful diversification away from the oil sector. A backdrop of low oil prices exacerbates diversification challenges but also provides impetus for change.

Saudi comprehensive reforms follow standard template to shift gears

The Saudi NTP broadly follows the generally accepted template of comprehensive macro- and micro- reforms needed to diversify beyond the hydrocarbon sector, in our view. The literature suggests successful diversification entails ideally a combination of three types of innovations: processes, products and organizations. Enhancement in processes would enhance productivity, new products would support emergence of new sectoral growth drivers, and improved micro- and macro-governance would help sustain production gains. The key to furthering the development path would be to move from an initial labour- and capital-intensive phase towards a phase focused on increasing productivity growth through higher value-added sectors.

This will require the retention of white-collar workers, steady progress on institution-building, an increase non-hydrocarbon FDI, horizontal and vertical integration, a broader manufacturing base including at first through sectors with competitive advantage (downstream or energy-intensive ones), greater integration into the global value chain through enhanced trade relations, as well as education and business climate reform to overcome structural rigidities, in our view.

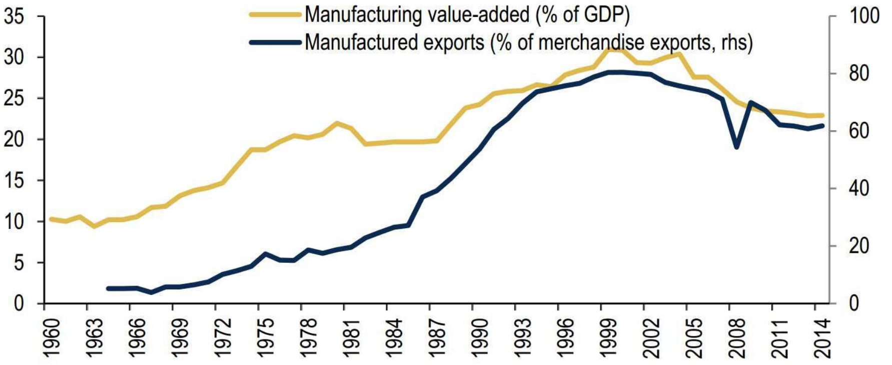
Medium-term diversification may require Fx reform

As diversification progresses, the case for increased Fx flexibility and making space for autonomous monetary policy conduct is likely to gradually take shape. Presumably, such a move would require putting in place a supportive institutional structure which is currently broadly lacking. It would also await improvements in productivity growth and the development of competitive local industries to minimize a potential Dutch disease effect on infant or other sectors. A fairly valued real effective exchange rate (REER) would support efficient allocation of production factors across sectors and improve competitiveness of the tradable goods sector.

Malaysia case study highlights the role of supportive public sector

Malaysia's diversification away from primary commodities relied on a series of National Industrial Policies and Industrial Master Plans to promote the manufacturing sector. Focus was given to sectors with high export potential, and efforts were taken to create linkages with other sectors and to deepen interconnection with other industries. Growth strategies also aimed to develop local technological capabilities and clusters of industrial development, similar to the localization and industrial strategies spearheaded by the Saudi National Transformation Plan.

Chart 27: The launch of government industrial policies in the 1980s helped Malaysia diversify



Source: Haver, BofA Merrill Lynch Global Research

Norway case study suggests institutions are paramount

While it is perilous to extend lessons that could overemphasize idiosyncracies, the case of Norway highlights, if anything, the importance of sound institutions and macro management when it comes to broader economic diversification. The case of Norway has relevance for the Gulf Cooperation Council countries (GCC), including Saudi Arabia, especially since oil discovery (late 1960s) and oil production (1970s) timelines were not that sensibly different from the GCC.

That being said, outcomes were widely different as well as the starting point since Norway was already a developed country with mature social, economic and political institutions at the time of the discovery of oil. This allowed a distinction between the management and ownership of natural resources uncommon in the GCC and several other resource-based economies, in our view.

Four ways in which the supply-side matters for economic diversification

We believe that there are four lessons to learn from Norway's outperformance:

1) The importance of human capital

Norway's priorities from early on were to build human capacity, investing in education, increasing labor force participation and supporting productivity growth.

2) Prudent conduct of fiscal policy

Current Norwegian oil revenue management puts considerable emphasis on stabilizing the economy and facilitating a gradual phase-in of oil revenues over time (which crowds out trophy projects and put onus on achieving productivity gains in the non-oil sector). Following the 1970-80s boom-bust sequence and expansion of the welfare state, Norway established in 1991 a Government Petroleum Fund to receive and invest oil revenue (it received its first net transfer in 1996). The fiscal rule adopted in 2001 targets a central government structural non-oil budget deficit equal to 4% of GPF assets (the latter assumed to be its estimated long-run real rate of return). That being said, Saudi Arabia's infrastructure requirements could have prevented adapting this part of Norway's model for long, in our view.

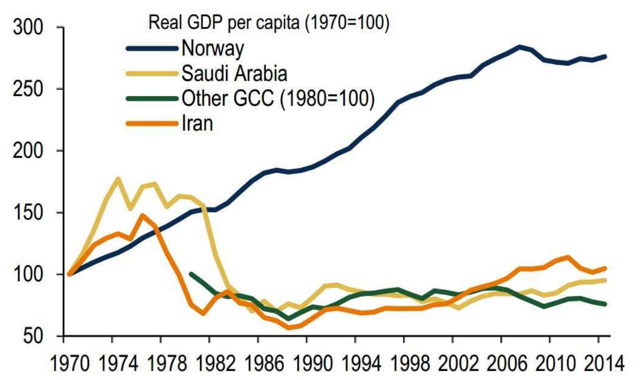
3) Institutional experience

Norway's development has been characterized by continuous development and integration of resource-based export-oriented industries, some of which were active since 1950. While Norway is the world's third largest exporter of natural gas and the sixth largest of crude oil, it is also the second largest exporter of seafood, possesses the fourth largest shipping fleet, is the sixth largest exporter of aluminium and the first exporter of sub-sea technology products and services.

4) Industrial policy

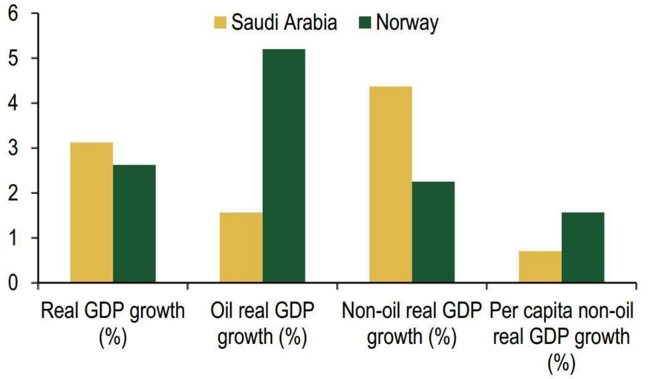
Norway's active use and introduction of a national industrial strategy has helped shelter infant industries and create linkages between the natural resource-based industries and other sectors of the economy. For two decades (1972-1994), preference was given in a transparent fashion to local content in procurement to help build supply industries, and provisions were set in for international firms to train nationals, use domestic service industries and cooperate in R&D with local institutions. Though Norway pre-dates the GCC in this development, the region shares in many ways some of these policies, leading to the emergence of national champions. The main difference appears to relate to the scope, extent and degree of innovation involved in offshoot industries at this stage, in our view.

Chart 28: Norway real GDP per capita growth outpaced GCC over the past decades



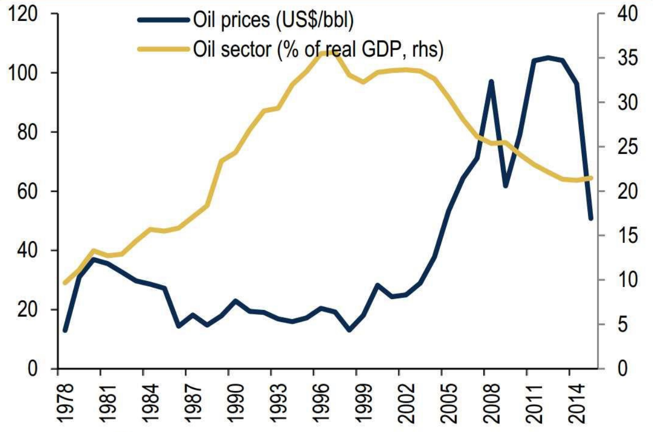
Source: Haver, BofA Merrill Lynch Global Research

Chart 29: Norway outperforms Saudi indicators on a per capita basis



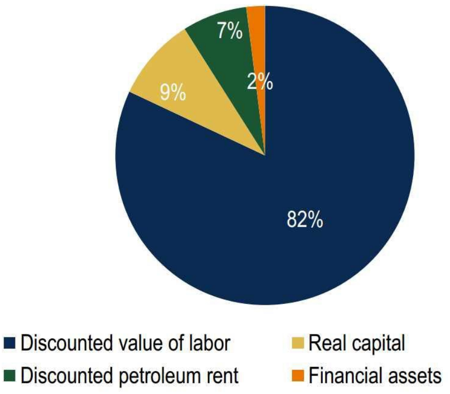
Source: Haver, BofA Merrill Lynch Global Research. Data represents averages over 1979-2014.

Chart 30: Norway's hydrocarbon sector share in real GDP has dropped



Source: Haver, BofA Merrill Lynch Global Research

Chart 31: Human capital is Norway's most important resource



Source: Norway Ministry of Finance, BofA Merrill Lynch Global Research

Eurobond premium required for fiscal slippage risk

We expect large and regular sovereign eurobond issuance going forward to support FX reserves and domestic liquidity but weigh on regional bond spreads if risk appetite does not hold up or fiscal slips. EMBIG index inclusion is unlikely, in our view. Saudi Arabia CDS premium to Qatar and likely larger issuance size suggests Saudi External Debt (EXD) is likely to be issued at a premium to Qatar.

Saudi Arabia international bond issuance latest in wave of Gulf supply

Local press suggests that Saudi Arabia is gearing up for a large international bond issuance program (US$10-15bn issuance target this year), which, if confirmed (the government neither confirmed nor denied , should add to the sizeable sovereign bond supply pipeline this year. So far, we have seen US$17.1bn in gross and net supply from the Gulf Cooperation Council (GCC) countries. Assuming Saudi Arabia issues US$15bn and accounting for planned issuance from Dubai and Bahrain and excluding Kuwait, the rest of this year should see a further US$16.5bn in gross issuance and US$16.1bn in net issuance. This would bring the total gross and net GCC sovereign bond issuance to US$33.6bn and US$33.2bn this year respectively.

Seminal potential bond issuance has multi-pronged implications

If confirmed, this could be a seminal event and mark the first time for the Saudi government to issue external bonds. It would allow for participation of foreign investors as domestic debt is being sold to Saudi banks and funds. External issuance should diversify funding sources, lock in still low interest rates, support SAMA's Fx reserves and support domestic liquidity; the latter two macro variables have weakened this year. External issuance would also imply the presence of an underlying asset that could theoretically trigger CDS contracts, as opposed to the current situation.

Investment grade status to be retained despite likely further rating cuts

A prolonged oil price downturn is likely to continue to put downward pressure on Saudi Arabia’s credit rating in the medium-term. On average, the GCC has benefited from three notch rating upgrades over the period 2002-10 prior to the start of the Arab Spring. Saudi Arabia was upgraded six times, from Baa3 to Aa3, by Moody's over the period starting from 1999, while it was upgraded two and three notches by S&P and Fitch over the period starting from 2003/2004 respectively. Moody’s has thus in the past preserved Saudi Arabia’s investment grade rating at the bottom of the cycle, as government debt to GDP stood at 103% of GDP and SAMA’s foreign assets at US$17bn (10% of GDP) in 1999.

EMBIG Index inclusion is unlikely

EMBIG index inclusion is unlikely, in our view. Estimated 2013 (US$25,140) and three-year rolling GNI per capita (US$23,090) data likely suggests Saudi Arabia does not meet EMBIG income index inclusion criteria, in our view. Saudi Arabia would however likely be eligible for Barclays EM Hard Currency Aggregate Index, based on its IMF classification as a non-advanced country (noting that the index provider has moved away from solely using rating for EM country classification). This suggests that Saudi policy-makers need to articulate a comprehensive, timely and credible medium-term fiscal policy to facilitate wide take-up from domestic, regional and international investors, in our view.

Pricing matters

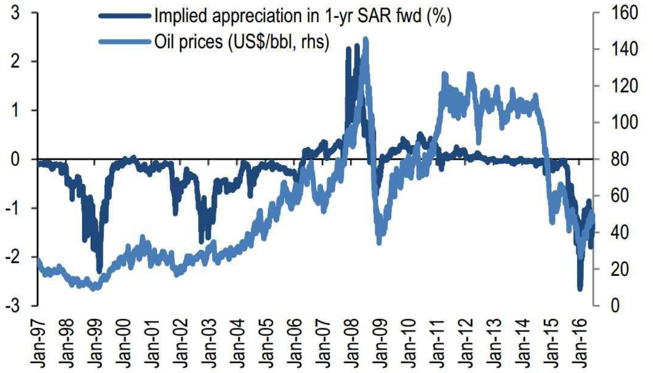
Saudi Arabia CDS premium to Qatar, relative credit metrics and likely larger planned issuance size suggests Saudi EXD is likely to be issued at a premium to Qatar, in our view. The closest regional peers to Saudi Arabia (A1/A-/AA-) are likely Qatar (Aa2/AA/AA) and Abu Dhabi (Aa2/AA/AA). Qatar’s existing external debt curve makes it a potentially useful and relevant pricing benchmark in this regard, in our view. The Qatar 10-year sovereign bond currently trades at yields of 3.0%. However, Saudi CDS has been trading c65bps wider of Qatari CDS, which would suggest a 10-year Saudi bond yield of c3.65%. Current CDS spread levels suggest potential for some notch downgrades from Saudi Arabia's current rating, as the market prices in issuance risk, volatile oil prices, hedging flows relative to the USD peg and the banking sector off-balance sheet wrong-way exposure risk.

Saudi forwards, rates and CDS remain under pressure

Domestic rates are under pressure with the 5y IRS spread over US at historically elevated levels due to structurally tighter liquidity. Budget consolidation could help take some of the pressure off as it would imply lower debt issuance needs. However, it would also imply lower deposit formation in domestic banks as fiscal retrenchment will impact private sector economic activity. Greater risk premium, issuance pressures and tighter liquidity will also keep CDS spreads elevated until oil recovers.

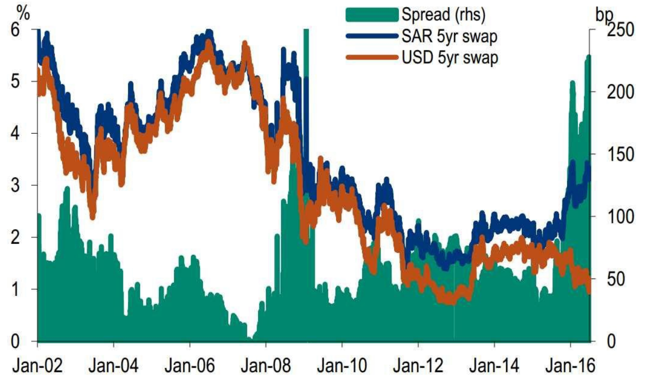
Hedging against SAR devaluation risks remains strategically attractive given the risk-reward ratio. Data suggest a strained backdrop for external and domestic liquidity reflecting the combination of private sector dollarization, possible capital outflows, weakening deposit formation and the move higher in interbank rates. However, SAMA has already intervened through macro-prudential tools and may do so again if needed.

Chart 32: Market remains nervous on Saudi Arabia



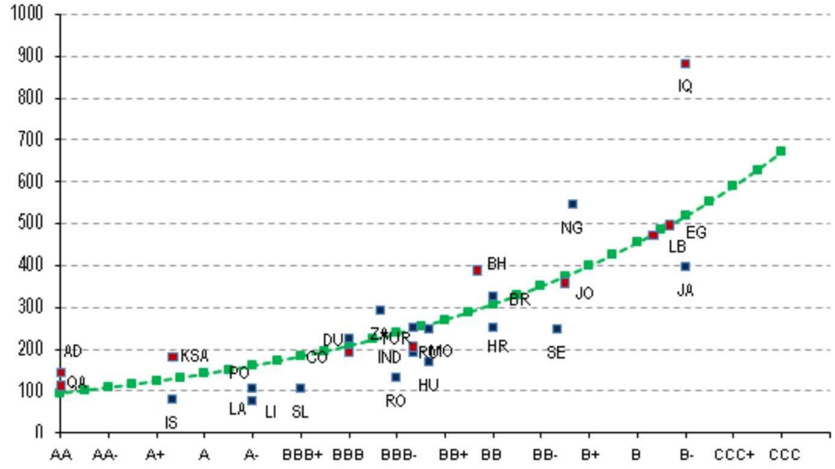
Source: Bloomberg BofA Merrill Lynch Global Research.

Chart 33: A pronounced increase in SAR swap spreads vs USD



Source: Bloomberg BofA Merrill Lynch Global Research.

Exhibit 2: Saudi credit risk prices in rating downgrades



Source: Bloomberg BofA Merrill Lynch Global Research.

Commodities: NTP adds to medium-term oil market tightness

Francisco Blanch
MLPF&S
francisco.blanch@baml.com

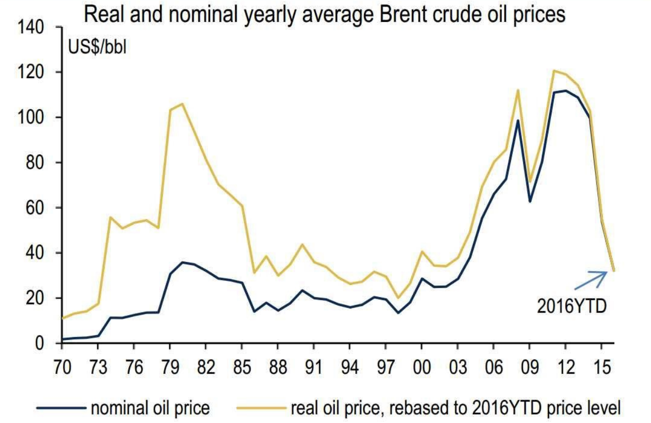
Peter Helles
MLI (UK)
peter.helles@baml.com

The Saudi National Transformation Plan (NTP) suggestion that production capacity is to be maintained at the current level until 2020 reinforces our conviction of medium-term oil market tightness. We continue to see a call on OPEC of 4.1mn bpd in the next five years to “balance the market”, and see oil prices of US$55-75/bbl over 2016-2020. Although the NTP incorporates a target of boosting domestic gas production by 50% to 18 bcf/d by 2020 which could crowd out oil demand, we expect Saudi Arabia’s ability to switch to reduce oil burn in power generation to 2020 to be quite limited, in our view.

We see oil averaging US$55-75/bbl over 2016-2020

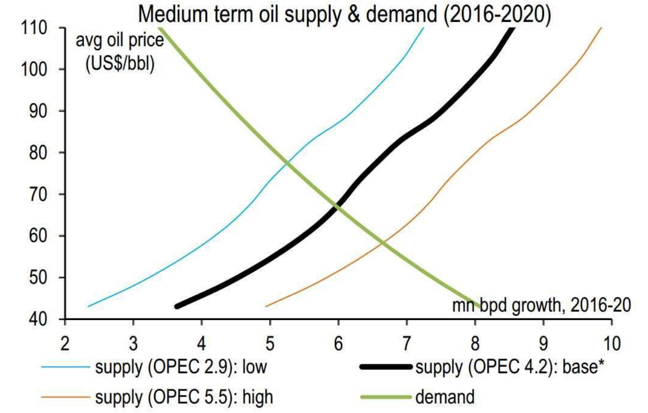
Real Brent prices are at one of the lowest levels in decades, and will likely encourage very strong demand growth ahead, while we see supply falling across non-cartelized producers. As such, we see global light sweet crude oil averaging US$55 to US$75/bbl over the 2016-2020 period, depending on how much incremental production OPEC can and will supply over the next five years. We think it likely that Saudi will dig into its untested spare capacity to go for increased market share, though uncertainty over how much Saudi can and will produce remains. We also remain concerned about output sustainability among weaker cartel members given the rapid credit profile deterioration.

Chart 34: Real Brent prices are at one of the lowest levels in decades, and will likely encourage very strong demand growth ahead



Source: BP, Bloomberg, BofA Merrill Lynch Commodities Research

Chart 35: We see global light sweet crude oil averaging US$55 to US$75/bbl over the 2016-2020 period

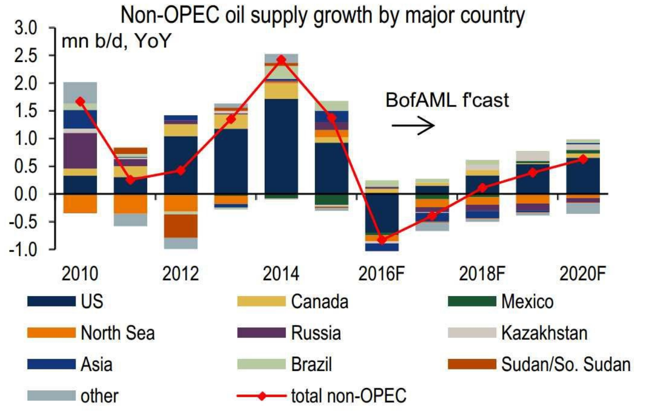


Source: IEA, BofA Merrill Lynch Commodities Research
*4.2 mn b/d OPEC supply: Saudi: 1.3 mn b/d; Iraq: 0.8 mn b/d; other OPEC crude: 1.7 mn b/d; 0.4 mn b/d OPEC NGLs.

Non-OPEC production will not reach 2015 levels before 2020 at the earliest…

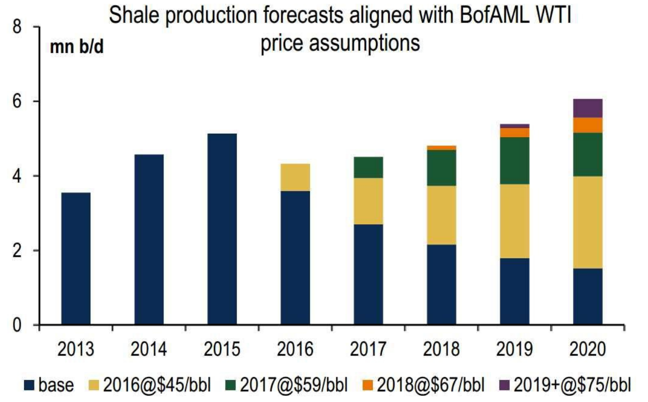
Non-OPEC producers have massively reduced capex spending, down US$290bn or 42% from 2014 to 2016, in response to the low price environment. Should capex start to increase again in 2017, the effect on non-OPEC non-shale production is unlikely to be felt before 2020 at the earliest. Most of the decline in the short term comes from non-conventional output in the US, as shale is very price sensitive within a 12-month horizon. With Brent prices set to increase from US$46/bbl this year to US$80/bbl in 2020 in our base case, we believe US shale production will grow again, albeit at a slower rate than in the past four years. Total non-OPEC supply is set to drop to 56.4mn bpd in 2017 before rebounding to 57.5 mn bpd in 2020, a similar level as in 2015.

Chart 36: With Brent prices set to increase to US$80/bbl in 2020, we believe US shale output will grow again, albeit at a slower rate



Source: IEA, BofA Merrill Lynch Commodities Research

Chart 37: Linking this with our 5-year price deck suggests marginal US shale output grows in 2017 and acceleration thereafter

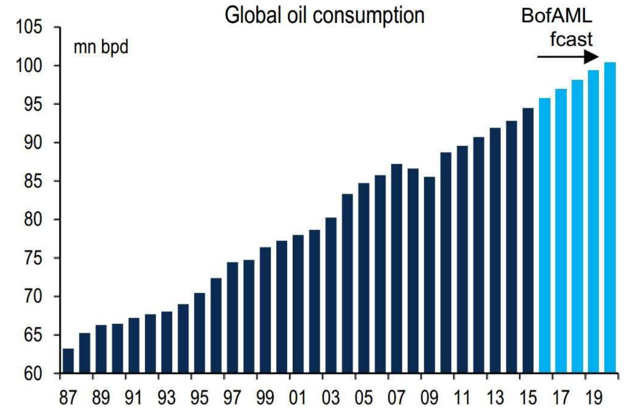


Source: EIA, BofA Merrill Lynch Commodities Research

4.1mn bpd needs to be added by OPEC by 2020, namely Saudi, Iran and Iraq

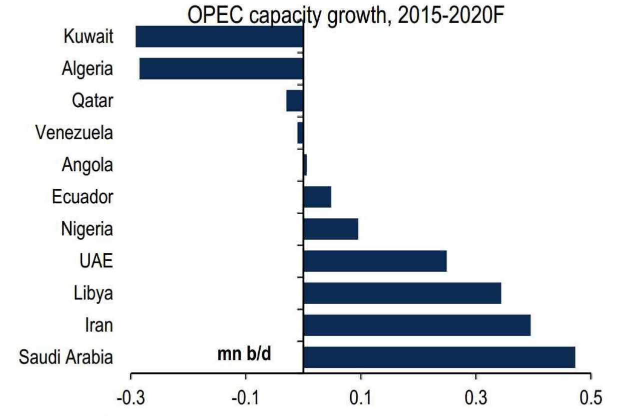
The US is the only country able to ramp up production among non-cartelized players by 2020, so OPEC has to come to the rescue to provide the required incremental supplies. We estimated that demand will grow by 5.9mn bpd in 2015-20. With the market oversupply of 1.8mn bpd in 2015, OPEC needs to increase production by 4.1mn bpd in the next five years to “balance the market”. Saudi Arabia could make up for half of this given its c2mn bpd of spare capacity, and we believe it intends to at least increase its market share. We would expect other OPEC countries to expand their capacity in the next five years, namely Iran and the UAE. As for Libya, the current turmoil needs to come to an end, while the scale of any Iraqi long-term output increase remains the biggest uncertainty.

Chart 38: We estimated that demand will grow by 5.9mn bpd in 2015-20



Source: IEA, BofA Merrill Lynch Commodities Research

Chart 39: Other than Saudi Arabia, some OPEC countries will also expand their capacity in the next five years, namely Iran the UAE

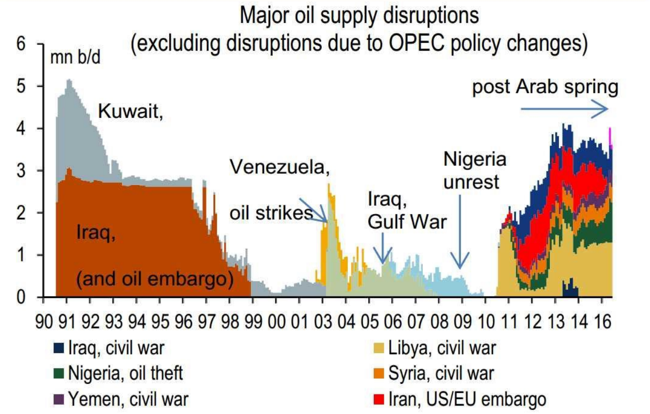


Source: IEA, BofA Merrill Lynch Commodities Research

Saudi plans to keep capacity flat add to medium term tightness

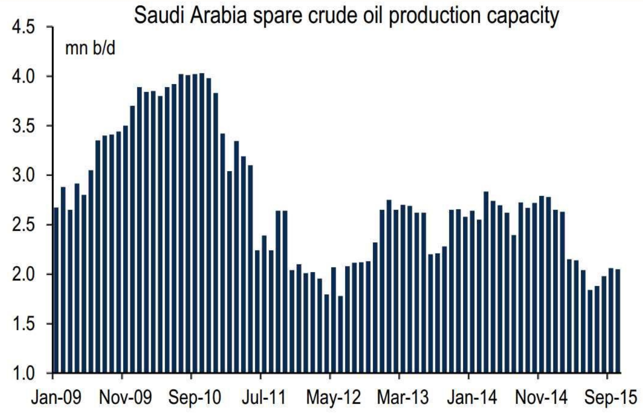
Saudi Arabia has 12.5mn bpd of crude production capacity, according to the government, and the country suggested within the NTP that it plans to maintain capacity at the current level to 2020. Saudi Arabia is the only material holder of spare crude oil production capacity around the world. Spare capacity currently sits at c2mn bpd, which is one of the lowest levels ever. Even if Saudi Arabia did start investing today, it would take a number of years to be completed and would unlikely be ready before 2020 in any case. Hence, as Saudi ramps up production over the medium-term by digging into its spare capacity, the risk premium in the oil market will rise as the market becomes increasingly less able to handle future supply disruptions.

Chart 40: The supply side in oil faces a lot of disruptions linked to geopolitics and broad economic mismanagement



Source: IEA, BofA Merrill Lynch Commodities Research

Chart 41: Saudi Arabia has 12.5mn bpd of crude production capacity, and currently is c2.0 of spare capacity over current production



Source: IEA, BofA ML Commodities Research. IEA estimates Saudi production capacity at 12.2mn bpd

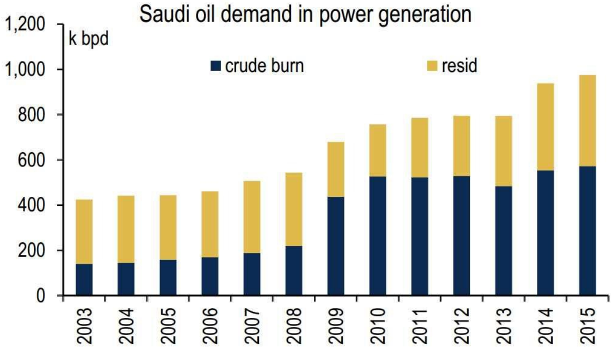
If Saudi gas production gets off the ground, it could crowd out oil demand...

The Saudi NTP plans to grow domestic gas production by 50% to 18 bcf/d by 2020, and to use most of this gas to meet a 30% rise in domestic power demand and moreover boost gas' share in power generation to 70%, up from 50% currently. This implies that oil burned for power generation could drop by 30% or 300 thousand bpd by 2020, most of which would likely be crude which is more valuable in the export market than the residual fuel oil burned. Saudi Arabia has the world's 4th largest gas reserves, yet reaching the 5.8 bcf/d production growth target by 2020 is not that straight forward.

…though boosting gas production to the NTP target poses significant challenges

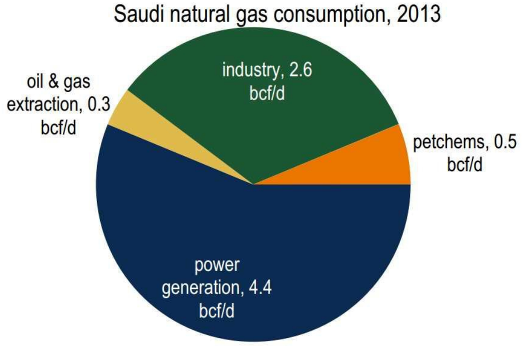
The new Wasit project, which is due to come online in 2016 and ramp up to 2.5 bcf/d by 2020, meets about 40% of the gas production growth target. Other than that, Saudi Arabia is focusing on getting unconventional production off the ground to meet the remaining 3.3 bcf/d of the production growth target. The timing, cost and commercial feasibly of which this unconventional gas is highly uncertain. The 2.5 bcf/d from the Wasit project is about the amount of gas that would be needed to meet incremental power demand and displace 300 thousand bpd of crude burn, assuming gas demand does not grow in other sectors, which is unlikely. Hence, it is likely that Saudi's ability to switch significantly to reduce oil burn in power generation to 2020 will be quite limited, explaining the new openness to consider gas imports, in our view.

Chart 42: Oil burned for power generation could drop by 30% or 300 thousand bpd by 2020, most of which would likely be crude



Source: IEA, BofA Merrill Lynch Commodities Research

Chart 43: Other sectors like petrochemicals and industry will see their gas demand grow as well



Source: IEA, BofA Merrill Lynch Commodities Research

Equity Strategy: more clarity required, but investible themes emerging

Hootan Yazhari, CFA >>
Merrill Lynch (DIFC)
hootan.yazhari@baml.com

More clarity required, but high level benefits becoming evident

The recent National Transformation Program (NTP) document provided a much awaited articulated roadmap through 2020 of how the government is seeking to achieve its highly ambitious Vision. However, we believe significantly more detail is still required before the market can make more concrete conclusions on the key winners and losers from the program as well as quantify the impact it could have on corporate earnings (both positive and negative).

Whilst at this juncture it is difficult to quantify the size of the NTP opportunity for individual companies, we feel more confident in highlighting the sectors that we see could benefit from Saudi Arabia ambitious plans. We have detailed the sectors that we see as likely benefiting most in the table below. We also highlight the sectors which could face the most negative impact from the roll out of the NTP.

Table 13: Summary of key sectors likely benefiting from and impacted by the National Transformation Program

Sectors likely benefiting from NTP	Rationale
Petrochemicals	With US$ 11bn being allocated to project development, downstream chemicals second largest focus of NTP; significant increase in natural gas availability (indigenous and imported); expansion opportunities
Healthcare	Volume growth opportunities for private healthcare providers (management contracts); private health insurance growing; participation in privatisations; significant increase in local pharmaceutical production
Insurance (healthcare)	Private insurance coverage to increase to 31mn from 10.5mn currently; rising availability of private healthcare facilities
Real estate	Higher home ownership targeted; improved access to financing/housing subsidies to both developers and buyers; volumes set to expand as NTP targets higher real estate sector growth; increased opportunities for private sector investment
Consumer staples	Demand growth from religious tourists; formalisation and Saudisation of retail sector; growth in locally produced poultry sales; longer-term private sector job creation
Telecom	Government spending US$2bn to enhance FTTH and Mobile networks; focus on increasing internet usage; religious tourism will boost demand
Metals and mining	Provision of new mining licenses; potential investment opportunities for international mining companies
Defence	Increased localization will cut imports, develop local industrial capabilities and create jobs

Sectors likely negatively impacted by NTP	Rationale
Consumer discretionary	Near-term squeeze on consumer disposable income and sentiment; increased competition as foreign entities given 100% ownership entitlement; development of Saudi postal system possibly a precursor to online shopping increase
Petrochemicals	Potential reduction in feedstock subsidies; potential for increased competition from international companies
Real estate	Land tax on white land could impact cost base and prices in urban areas; higher competition from new entrants
All sectors	Rising costs on reduced subsidies (particularly water and energy); higher wage costs; increased financing burden on private sector as government partially shoulders NTP costs

Source: National Transformation Program, BofA Merrill Lynch Global Research,

Six key investible themes from the NTP are emerging

Using this frame work we believe a number of investible themes are emerging including:

1. **Ambitious plans to grow religious tourism:** with Hajj visitors set to increase from 1.5mn to 2.5mn and Foreign Umrah Pilgrims from 6mn to 15mn per annum, we see significant potential for the travel & tourism, transport sectors, consumer discretionary and telecom services sectors to benefit.

2. **Down trading as pressure mounts on the consumer:** With the NTP looking to reduce subsidies on energy (including transport fuels) and water, we see pressure on the disposable income of Saudi consumers rising in the short- to medium-term (although in the longer term, we expect higher employment levels, growing home ownership and accelerating growth to offset these factors). These issues could be further augmented by a slowdown in public sector wage growth and growth in the proportion of residents employed by the private sector,

   In reflection, we believe that the consumer will likely become more value conscious in the near- to medium-term as stress on consumer discretionary spend mounts. As such, we continue to prefer the consumer staples and the grocery retailers in this environment. We also believe the grocery retailers will benefit from the government's focus on formalising the sector (it is highly fragmented currently and dominated by smaller independent stores) in an effort to increase Saudi participation. That said, we believe consumer discretionary companies which offer more affordable and economic goods could benefit as they take market share.

3. **Rising focus on healthcare provision:** Saudi Arabia is focused on significantly growing accessibility to healthcare for its residents. The provisions in the NTP include a significant increase in medical centres (hospitals and clinics), wider provision of private health insurance and an increase in the level of

pharmaceuticals manufactured locally. We thus see potentially strong opportunities for the private hospital operators, the health insurance providers and local pharmaceutical manufacturers.

4. **An improvement in the availability of affordable housing for Saudi nationals:** The Saudi housing shortage has been a persistent issue for the country; with the shortfall currently standing at an estimated 1.2mn units, in our view. Looking forward, we believe this shortfall will likely intensify given strong demand formation. The NTP is seeking to introduce greater numbers of affordable housing units and greater availability of financing. Indeed, the ministry of housing is looking to invest cUS$16bn in achieving these aims by 2020. Whilst we do not believe these measures go far enough to fully resolve the housing shortage, we nevertheless see material opportunities for Saudi real estate developers.

5. **Growth in telecom/fibre infrastructure:** The NTP is seeking to greatly enhance connectivity in the country, largely through increased usage of internet. Reflecting that, the government has pledged approximately US$2bnn to significantly increase the availability of high speed FTTH networks in remote areas, increase mobile broadband coverage and density in urban centres (3G/4G) and allocate an increased amount of bandwidth to the mobile service providers. We thus see significant volume growth potential for the telecom service providers (particularly with regards to mobile data, where we see scope for price increases). Furthermore, we expect much of the capital expenditure required to expand the FTTH and Mobile networks to be provided by the private sector, which in our view will accelerate the case to spin off their tower portfolios (as a method of funding capex expansions).

6. **Significant growth in downstream petrochemicals capacity and metals & mining:** The NTP's drive to diversify revenues away from the oil sector has seen it focusing on exports of non-oil commodities. In particular, we highlight the significant investment it has earmarked for the downstream industries including petrochemicals and oil refining, as well as aggressively growing the country's metal and mining operations (likely through national champion, Maaden). Examples include the expansion of the country's base metals production (e.g. bauxite and phosphates) and the construction of an oils-to-olefins project. The further development of capital intensive industries also provides some scope for Saudi Arabia to develop service industries and increase local content provisions, in our view. These could be facilitated through joint-ventures with global service providers, who may consider manufacturing plants in Saudi Arabia.

How to gain exposure to NTP themes

In reflection of the themes articulated above (and in the sections that follow), we highlight the sectors and companies which have exposure to each of these themes in the table below. The companies highlighted in green represent the companies which are currently Buy rated by our fundamental analysts and also fit in with our theme of gaining exposure to the NTP.

Chart 44: Summary of sectors and companies with exposure to key NTP themes (companies highlighted in green are rated Buy and represent our top picks)

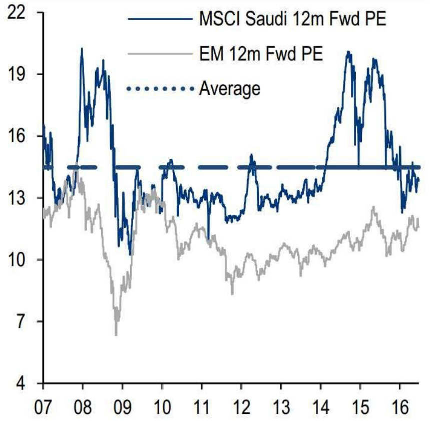
	Theme 1	Theme 2	Theme 3	Theme 4	Theme 5	Theme 6
	Religious Tourism	**Downtrading**	**Healthcare provision**	**Growth in Affordable housing**	**Telco infrastructure growth**	**Increase in none-oil commodity exports**
Sectors to benefit	Travel & Tourism	Consumer staples	Private Hospitals	Real Estate	Telecoms	Petrchemicals
	Transport	Consumer Disc.	Insurance co's	Construction	Training & Education	Metals & Mining
	Consumer staples	Training & Education	Pharma	Banks		Industrials
	Training & Education		Training & Education	Training & Education		Training & Education
BofA ML coverage with exposure	**STC**	**Savola**	**Al Hammadi**	Dar al Arkan	**STC**	**SABIC**
	ZAIN KSA	Almarai	Dallah		**Zain KSA**	Maaden
	Savola	**Al Othaim**				**Yansab**
	Almarai	**Al Hokair**				Safco
	Al Othaim	Jarir				
Other Companies with exposure	Mobily	Sadafco	Mouwasat	Emaar Economic city	Mobily	Petrorabigh
	Sadafco	Halwani Brothers	Care	Saudi Banks	Al Khaleej Training	Zamil
	Al Tayyar	Nadec	MEAHCO	Al Khaleej Training		Al Khaleej Training
	Dur Hospitality	Saudi Marketing	Spimaco			
	Saudi Ground Serv.	Al Khaleej Training	Bupa			
	Saudi Catering		Tawuniya			
	Al Khaleej Training		Med Gulf			

Source: BofA Merrill Lynch Global Research

Saudi market valuation not testing

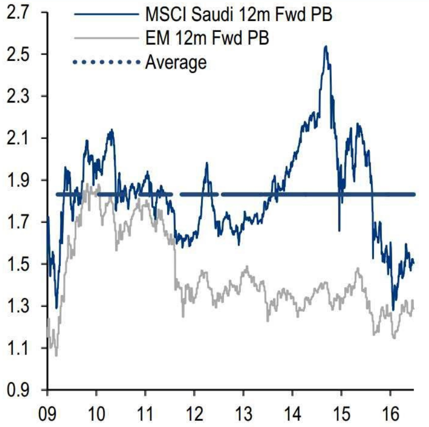
The Saudi market has been one of the poorest performing emerging markets year-to-date, underperforming the MSCI EM indices by approximately 10% (and c20% since August 2015). A combination of weak oil prices and their inevitable impact on the Saudi economy including slowing economic growth, rising operational risks and fiscal consolidation (including subsidy removals) have been the key drivers. The underperformance has left the market trading on a 12mth P/E of c.13x, an 11% discount to its long term average and its lowest premium to GEMs since 2010 (Saudi trades on a 7% percent premium vs long run average of c. 30%). Furthermore, Saudi trades on 12mth FWD P/B ratio of 1.5x, a 20% discount to long run averages, but a 15% premium to GEMs on 1.3x, justified by the market's higher ROE's (Saudi's long term P/B premium has been c.25%). Finally, Saudi trades on 12mth FWD EV/EBITDA multiples of 9x, a c.11% discount to long run averages.

Chart 45: Saudi 12m fwd. P/E vs. EM



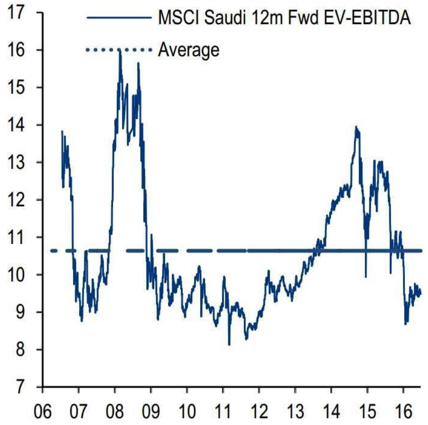
Source: DataStream, I/B/E/S, Bloomberg, BofA Merrill Lynch Global Research

Chart 46: Saudi 12m fwd. P/B vs. EM



Source: DataStream, I/B/E/S, Bloomberg, BofA Merrill Lynch Global Research

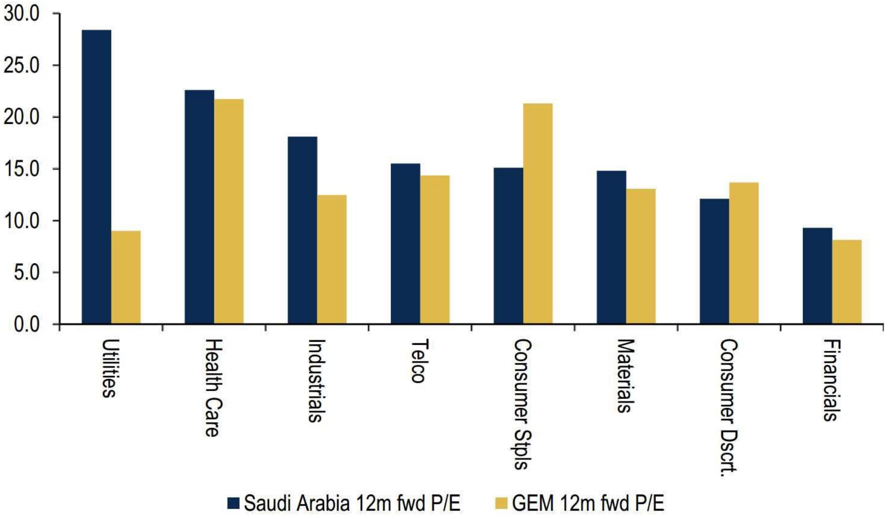
Chart 47: Saudi 12m fwd. EV/EBITDA



Source: DataStream, I/B/E/S, Bloomberg, BofA Merrill Lynch Global Research

Interestingly, not all sectors within the Saudi market trade at a premium to GEM peers, with both consumer discretionary and consumer staples trading at a material discount. The remaining sectors all trade at a modest premium to their GEM counterparts, with the industrial and utility sectors trading at a substantial premium on 12mth FWD P/E

Chart 48: Consumer sectors trade at a discount relative to GEM averages. Industrials and Utilities are relatively expensive vs. GEM peers, other sectors trade at small premium

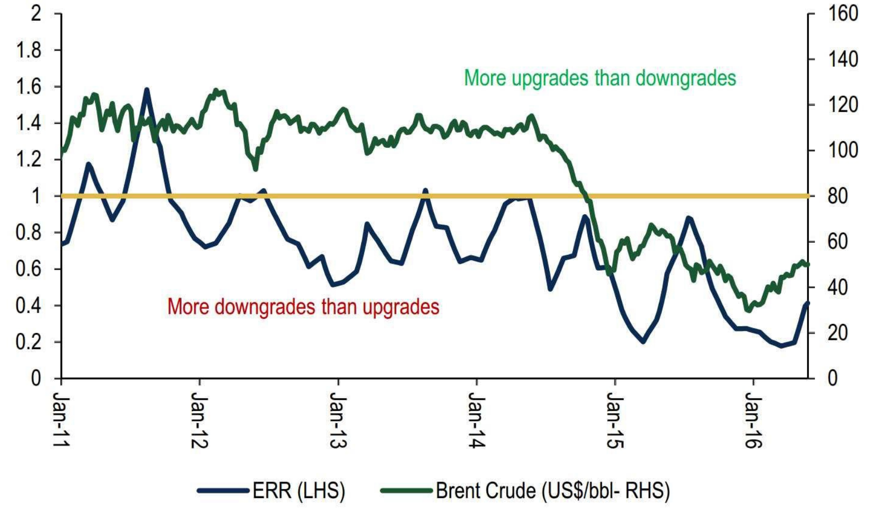


Source: DataStream, I/B/E/S, Bloomberg, BofA Merrill Lynch Global Research

Weaker economic outlook largely priced in as ERR shows inflection

A further consequence of the weakening economic outlook for Saudi Arabia has been the collapse in market expectations, with earnings estimates having posted their longest sustained period of downgrades on record. We note however, the Saudi Earnings Revision Ratio (ERR) has reached an inflection point, with a slowdown in the rate of earnings downgrades. (ERR is a ratio of number of companies enjoying earnings upgrades vs number of companies suffering downgrades. A ratio above one, therefore, indicates more upgrades than downgrades). Consequently, we believe the relatively un-testing valuations of the Saudi market indicate that a high level of negativity is already being priced in.

Chart 49: Saudi ERR's are showing signs of inflecting despite one of the most protracted periods of earnings downgrades.

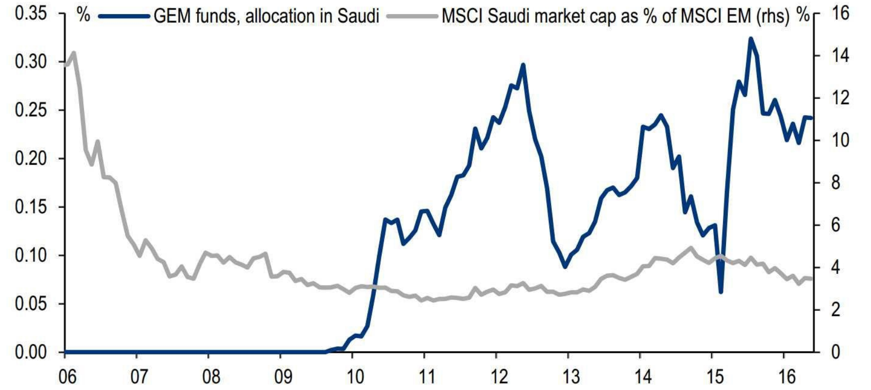


Source: DataStream, I/B/E/S, Bloomberg, BofA Merrill Lynch Global Research

Off benchmark and out of mind? Growing confidence could drive rerating

With Saudi Arabia remaining off benchmark (it is not currently part of any major indices such as the MSCI EM or FM), GEM funds allocation to Saudi Arabia remains relatively low at just 0.25% (of total GEMs AUM). Whilst allocation levels are relatively high versus history, they are still significantly below the levels implied by MSCI EM inclusion (1.4%) or suggesting a level of apathy towards the market.

Chart 50: GEM fund allocation in Saudi (asset-weighted)*



Source: EPFR, BofA Merrill Lynch Global Research.*Market cap as % of MSCI EM represents and artificially constructed benchmark weight assuming that this market was a member of MSCI EM index

With the NTP looking to diversify the Saudi economy away from the oil sector (and reduce its reliance on a higher oil prices), we believe confidence in the longer term economic outlook could improve, ultimately driving earnings momentum and a market rerating. Given our view that crude oil prices will likely post growing momentum in 2017, moving substantially above US$50/bbl, even a partial success of the NTP program would likely be sufficient to improve market confidence.

CMA reforms (likely linked to the NTP), a potential catalyst for market rerating

Furthermore, reforms being introduced by the Saudi Capital Markets Authority (CMA) to attract greater foreign participation in the Saudi market (see below) indicate Saudi is seeking an expedited path to inclusion within the major Stock indices. Given the NTP is seeking heavy involvement/investment from the private sector, we believe the measures introduced by the CMA to attract direct capital inflows to the country are likely linked with the larger NTP process. Growing confidence in the CMA's market reforms and the potential for inclusion in the major indices could prove to be a meaningful basis for the market to rerate upwards.

Taking positive steps to accelerate MSCI inclusion

The Saudi Capital Markets Authority recently announced extensive changes to the Qualified Financial Investor (QFI) program, which should ultimately increase the ease of foreign access to the Saudi market. In particular:

(1) Individual foreign investors will now be allowed to own up to 10% of the equity in a company (up from 5%). The total limit on foreign ownership remains at 49%.

(2) CMA decreased the minimum AUM of QFIs to US$1bn from US$5bn (although funds with AUM's below this level will be considered by the CMA).

(3) CMA introduced stock lending & covered short selling;

(4) CMA extended the settlement cycle to T+2 from T+0, effectively eliminating pre-funding;

(5) CMA effectively removed the concept of a QFI client, which would have prevented investors from using multiple fund managers to gain exposure to Saudi Arabia. This highly restrictive clause was a significant reason why many global institutions did not seek to gain QFI status; and

(6) CMA significantly reduced the amount of bureaucracy required to apply for and retain QFI status.

MSCI inclusion one step closer as accessibility increases

In our view, the range of measures the CMA is looking to adopt greatly enhances the accessibility of the Saudi Stock exchange and as such bring Saudi Arabia one step closer to inclusion in the MSCI EM index. Whilst we continue to see inclusion most likely from 2019 (with an announcement in 2018), the new changes indicate that a 2018 inclusion (with an announcement in 2017) is by no means out of the question. Indeed, whilst the timing of introduction of these new measures is likely to be 1H2017, a recent release of the draft proposals from the CMA suggested it could come as soon as 3Q16; although we await further clarification.

A higher weighting in international indices

Saudi Arabia is currently under consideration for inclusion in the MSCI EM index, potentially as soon as May/June 2018; although we believe 2019 is more likely. Saudi Arabia is currently not included in any of the Major global indices and as such its inclusion would have a profound effect on the market as it:

(1) leads to significant inflows of capital in to the market from both active and passive funds following the MSCI EM index and;

(2) significantly increases interest in the Saudi market as it gains profile and is no longer seen as off benchmark.

We see these factors as compatible with the NTP aim to boost capital markets.

Currently Saudi would be 1.4% of MSCI EM = US$10.9bn of inflows.

As per MSCI's guidance, we use MSCI's standalone Saudi market index as a template (released May 12th 2015) for Saudi stocks to be included in the MSCI EM index. The MSCI Saudi index currently contains 19 stocks (all listed entities on the Tadawul market) which are likely for inclusion in the MSCI EM index (were Saudi to be included). By taking the total market cap of these companies and multiplying them by the foreign inclusion limit articulated by the MSCI (currently 0.2), we are able to decipher the relative weighting of Saudi Arabia if it were to be included the MSCI EM index.

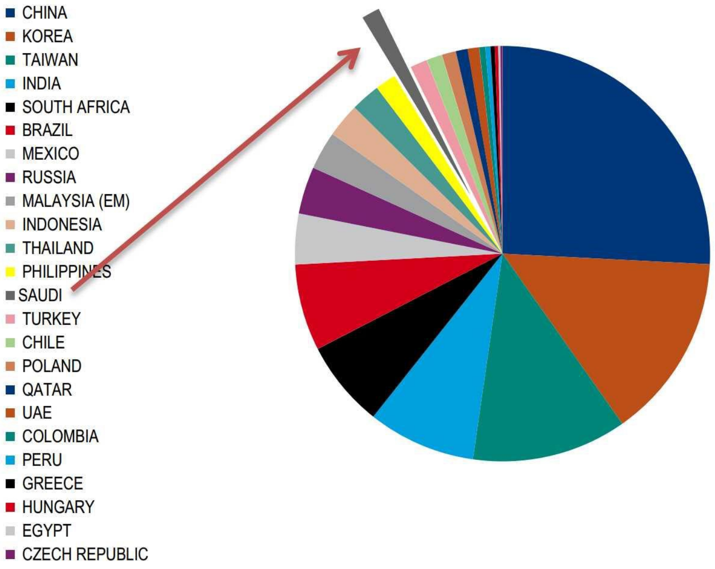
This analysis suggests that Saudi's potential weighting in the MSCI EM Index (were it to be included today) would be c1.4%. Furthermore, it would likely trigger net inflows (from other MSCI EM markets) of US$10.9bn when including all passive and benchmark-neutral active funds tracking MSCI EM. Importantly, our analysis is based on EPFR data, which respectively indicates US$405bn and US$354bn of active and passive funds following the MSCI EM index.

Table 14: Provisional MSCI Saudi Arabia constituents stand to gain as much as $10.9bn of total passive and benchmark-neutral active funds on inclusion

	Provisional weight in MSCI Saudi	Provisional weight in MSCI EM	ADTV (US$m)	Potential active buying (US$m)	Potential active days buying	Potential passive buying (US$m)	Potential passive days buying	Total US$m active + passive days buying	Total days active + passive days buying
AL RAJHI BANK	9.5%	0.1%	37.2	546.5	14.7	477.7	12.8	1,024	27.5
ALINMA BANK	2.0%	0.0%	223.0	116.9	0.5	102.2	0.5	219	1.0
ALMARAI	4.4%	0.1%	7.7	255.8	33.1	223.6	28.9	479	62.0
ARAB NATIONAL BANK	2.0%	0.0%	1.4	113.6	81.7	99.3	71.4	213	153.2
BANQUE SAUDI FRANSI	2.8%	0.0%	1.3	163.1	122.7	142.6	107.2	306	229.9
ETIHAD ETISALAT CO.	2.2%	0.0%	11.9	126.1	10.6	110.2	9.2	236	19.8
RABIGH REFN.& PETROCH.	1.0%	0.0%	9.2	59.1	6.5	51.7	5.6	111	12.1
RIYAD BANK	3.4%	0.0%	2.0	195.6	100.3	170.9	87.7	367	188.0
SAMBA FINANCIAL GROUP	4.0%	0.1%	5.0	232.9	47.0	203.5	41.1	436	88.2
SAUDI ARABIA FRTZ.	2.6%	0.0%	4.4	148.2	33.5	129.5	29.2	278	62.7
SAUDI ARABIAN MINING	4.5%	0.1%	17.2	260.8	15.1	228.0	13.2	489	28.3
SAUDI BASIC INDUSTRIES	25.3%	0.4%	170.9	1459.4	8.5	1275.6	7.5	2,735	16.0
SAUDI ELECTRICITY	8.4%	0.1%	6.3	485.3	77.3	424.2	67.6	910	144.8
SAUDI TELECOM	13.4%	0.2%	9.5	774.7	82.0	677.1	71.7	1,452	153.6
SAVOLA GROUP	2.0%	0.0%	6.9	114.4	16.6	100.0	14.5	214	31.2
YANBU NAT.PETROCH.	2.3%	0.0%	6.7	134.5	20.2	117.6	17.7	252	37.9
FAWAZ ABDULAZIZ ALHOKAIR	1.0%	0.0%	5.7	55.5	9.7	48.5	8.5	104	18.1
ALTAYYAR	0.8%	0.0%	33.5	44.4	1.3	38.8	1.2	83	2.5
NATIONAL COMMERCIAL BANK	8.2%	0.1%	6.8	474.8	70.1	415.0	61.3	890	131.4
TOTAL				**5,762**		**5,036**		**10,798**	

Source: BofA Merrill Lynch Global Research, MSCI and DataStream

Chart 51: If included today, Saudi Arabia would be the thirteenth largest constituent of the MSCI EM, accounting for 1.4% of the index.



Source: : BofA Merrill Lynch Global Research, MSCI and DataStream

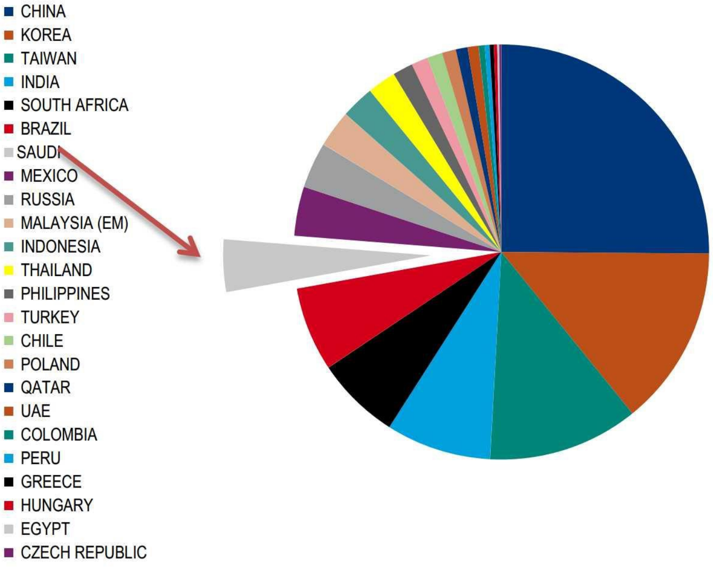
Privatisations & ownership limits could increase weighting

However, our analysis of Saudi Arabia's weighting could be significantly understated for two reasons, including: (1) our analysis currently uses a foreign ownership limit (FOL) of 0.2 (as previously guided by the MSCI). This could be understating Saudi's weighting given the increase in foreign ownership introduced in the CMA's announcement. Indeed, if we increased our FOL factor to 0.4, Saudi would be 2.8% of the MSCI and could attract US$21.3.bn of inflows; and, (2) Saudi Arabia (according to local press and the NTP) is seeking a number of privatisations in the coming 24 months, including a potential IPO of Aramco. Inclusions of these companies would likely increase Saudi's weighting in the MSCI EM index and thus attract higher inflows to the market.

Inclusion of energy assets could see Saudi accounting for 4.1% of MSCI EM

By way of example, we believe inclusion of Saudi Aramco in the Saudi market would profoundly affect the weighting of Saudi Arabia in the MSCI EM in our view. Indeed, if we were to replicate the analysis above using the Deputy Crown Prince hypothetical valuation of US$2tn for Saudi Aramco and a 5% inclusion factor (given 5% or less of Saudi Aramco would be listed according to interviews with the Deputy Crown Prince), Saudi Arabia would account for 4.1% of the MSCI EM index and likely attract cUS$31bn of inflows from passive and active funds. We note, our calculations are highly sensitive to the inclusion factor that MSCI would ultimately use (we assume 5%, in line with its estimated free float of 5%). An inclusion factor of 10% for example, would see Saudi Arabia accounting for 6.6% of the MSCI EM index.

Chart 52: If energy assets were included, Saudi Arabia would be the seventh largest constituent of the MSCI EM, accounting for 4.1% of the index



Source: BofA Merrill Lynch Global Research, MSCI and DataStream

Telecom: supporting a move to higher connectivity

Hootan Yazhari, CFA >>
Merrill Lynch (DIFC)
hootan.yazhari@baml.com

Table 15: Key NTP objectives for the Ministry of Communications and Information Technology

No.	Strategic objective	Key Performance Indicator (KPI)	Unit	Baseline	2020 target	Regional b'mark	Int'l b'mark
1	Provide critical resources, especially frequency spectrum for Information telecommunications and Technology services	Percentage of frequency spectrum available for telecommunication services out of the total allocated telecommunication services	%	42	80	NA	>90
2	Provide broadband services to all KSA regions by stimulating investment in infrastructure and developing tools, technical and regulatory frameworks	Percentage of FTTH coverage in densely populated urban areas	%	44	80	>95	>90
		Percentage of FTTH coverage in urban areas	%	12	55	>90	>80
		Percentage of wireless broadband networks' coverage (more than 10 Mbps) in remote areas	%	12	70	NA	74
3	Develop and activate smart government transactions based on a common infrastructure	Maturity level of the government services transformation to e-services	%	44	85	NA	NA
		KSA's rank in the United Nations index for the development of e-government	%	36	25	18	11
4	Bridge the digital gap in the skills of ICT users	Percentage of internet users in KSA	%	63.7	85	90.4	87.9

Source: National Transformation Plan

NTP helps increase penetration of higher margin offerings

We believe the National Transformation Plan will herald significant change for the Saudi telecom sector over the next five years. In particular, its focus on increasing access to high speed internet (via wireless and fibre) should underpin a material uplift in data usage and broadband penetration levels. Given these offerings represent the highest margin services provided by the telecom service providers, we believe the move towards a more data intensive society should be supportive for industry margins. Furthermore, with the NTP's plans relying heavily on private sector investment (including STC), we believe it could accelerate the spin out of tower portfolios from the three telecom players in the Kingdom (as a way of raising financing).

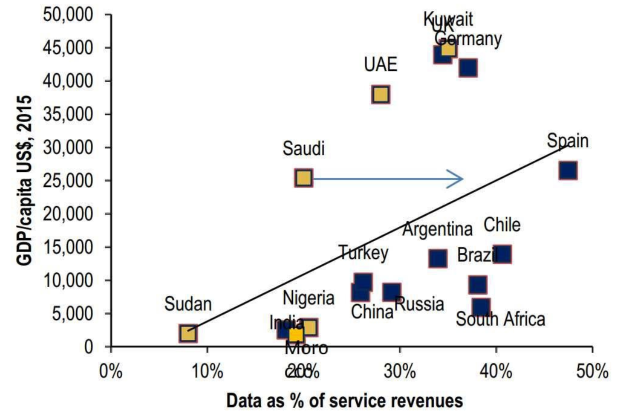
Government investment set to exceed US$2bn

As part of the National transformation plan, the government is seeking to invest more than US$2bn in the expansion of wireless (3G & 4G) and fibre infrastructure (FTTH) across the kingdom. Whilst the finer details on how this expenditure will be deployed (e.g. via grants or via direct investments?) have yet to be disclosed, details of the NTP suggest it will be used to expand connectivity in the remote areas of Saudi Arabia (given the telecom companies have largely avoided such expenditure in the past given low returns), as well as increasing network density in developed and urban areas.

Spectrum increases an indication of ambitious growth

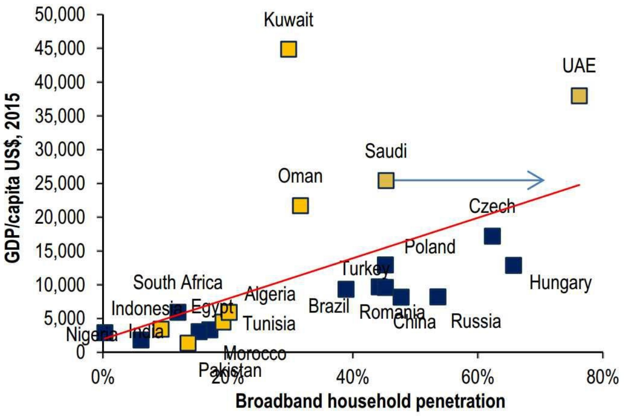
The government is also looking to significantly enhance the availability of spectrum to the telecom industry (from 42% to 80% of total spectrum allocated), highlighting the expected surge in data usage in the Kingdom to 2020 and beyond. Whilst data usage in the Kingdom has been rapidly growing in recent years on the back of high speed rollouts (3G & LTE), demand still lags global and regional averages (as indicated by data as a percentage of total mobile revenues). Ultimately, we believe this will have a number of effects including: (1) Saudi Arabia will further increase its focus on high speed offerings, (LTE/LTEA and eventually 5G); (2) capex cycles will likely lengthen as telecom companies intensify their rollout of coverage; and, (3) tower density will likely have to increase across the key demand centres.

Chart 53: Data as % of mobile service revenues



Source: BofA Merrill Lynch Global Research, Telegeography and company data

Chart 54: Household Broadband penetration



Source: BofA Merrill Lynch Global Research, Telegeography and company data

Private sector participation and investment will be crucial…

Whilst the government has earmarked cUS$2bn for the expansion of FTTH and wireless infrastructure (with a view to ultimately increasing internet usage), we believe this will likely be insufficient to meet the needs of the NTP's ambitious targets (based on costs of rolling out existing wireless and FTTH networks). As such, we believe the private sector (Zain KSA, Mobily and STC) will be key contributors to the financing of the expansion plan.

…intensifying the case for creation of a Saudi tower company

The Saudi telecom service providers have invested heavily in rolling out high speed wireless networks and, in the case of Mobily and STC, FTTH networks (not to mention license costs). This has seen the balance sheets of both Mobily and Zain KSA reaching relatively high gearing levels; thereby limiting their ability to step up capital expenditure for a sustained period. Consequently, we believe the focus on raising capital (to fund the expansion programmes) from the spin out of their Tower portfolios will only increase. We see **Zain KSA** as the key beneficiary of this given they have the most highly geared balance sheet amongst Saudi Telco peers.

Mobile market share gains for Zain KSA and Mobily increasingly important

Whilst the Saudi government has taken steps to introduce competition in the telecommunications industry, its moves to open the market have thus fallen short of this objective. Specifically, government-owned STC retains more than 60% revenue market share in Saudi Arabia, giving it a dominant position in the market, whilst Mobily and Zain KSA have only managed to achieve c25% and 15% revenue market share respectively.

This, in our view, has likely been driven by a number of factors including: (1) the relatively high cost of market entry (ie licenses) for both Mobily and Zain KSA, which has lumbered both with high operating costs; (2) high royalty costs, which amount to c16% of revenues generated in Saudi Arabia (with some exceptions, including data), arguably inhibiting requisite marketing and infrastructure spend. We note both Mobily and Zain KSA remain loss making; and, (3) the relatively high mobile termination rates, which arguably afford an advantage to STC and prevent Zain or Mobily from competing more aggressively on price.

For the realisation of the NTP, we believe it is important that all private sector players are realising sufficient rates of return and FCF generation to be able to finance the requisite growth in infrastructure. We note both Mobily and Zain KSA were loss making in 2015 and are expected to return sub 5% ROE's in both 2016 and 2017 (according to

Bloomberg consensus estimates) vs. STC at 17% (consensus). It is thus arguably crucial for Zain and Mobily to be allowed to increase market share in the mobile arena. We thus believe the regulator (CITC) regulator will have to consider further steps to afford market share to the second and third entrants including asymmetric competition measures (pricing, MTRs), license extensions or differentiated royalty rates.

STC the likely beneficiary of FTTH expansion

Whilst we argue that STC could lose further market share in the Mobile market as a result of regulatory moves, we believe the negative impact will be offset by growing revenues from FTTH, IPTV and the corporate market. Indeed, we believe STC, given its market leading FTTH network and IPTV offering, will likely benefit most acutely from the growth in the number of FTTH customers. Furthermore, with government grants likely to be provided for FTTH roll out in less economical areas (as part of the NTP), we believe returns will likely not suffer from the additional capital deployment. Mobily could also benefit (given its well-developed FTTH infrastructure), although we expect it to lag STC who has a first mover advantage and arguably a stronger FTTH/Triple play offering.

We note, Zain KSA will unlikely participate in the rollout of the FTTH expansion given its constrained balance sheet and strategic focus on the wireless market. That said, the company could benefit from the rollout of high speed internet solutions over its wireless network (ie fixed broadband over its 4G network), particularly in areas where the construction of FTTH networks maybe be difficult or uneconomical.

Royalty rate increases unlikely, particularly for Zain KSA and Mobily

Admittedly, an increase in royalties could be an easy way of raising much needed revenues for the government. After all, each 1% increase could raise income by more approximately SAR500mn (cUS$130mn). However, we would argue that this is unlikely given the lack of sufficient competition in the market and the obstacles to increased competition higher royalty rates would introduce. In the same vein, we also believe it highly unlikely that the government would seek to introduce a fourth mobile operator license given the lack of sufficient market capacity between the second and third operators.

Religious tourism a material opportunity for telecom providers

Saudi Arabia's ambitious growth targets for religious tourism provide a material opportunity for the telecom providers in our view. More specifically, if the number of religious tourists (for Hajj and Umrah) increases by an estimated 10mn per annum by 2020, we believe demand for roaming services (voice and data) and sim purchases will increase significantly. Indeed, on our estimates, we see the 10mn visitors potentially providing a SAR2bn-SAR4bn opportunity, representing a c4-8% uplift in total wireless revenues for service providers. We particularly believe **Zain KSA** will benefit strongly from this given they provide the most competitive Pay as you go packages currently and have the largest spare capacity on their network.

Health: Not Tremendously Prescriptive

Vision 2030 a start but Needs Transparent Proposals

The Saudi government's Vision 2030 document calls for an improvement in healthcare in Saudi Arabia. The private healthcare sector has some role to play; the National Transformation Plan (NTP) targets the private sector to be responsible for funding 35% of healthcare spend in 2020, up from 25% today. How it intends to do this is uncertain. We believe the incumbent private hospital operators, including listed companies (Al Hammadi, Care, Dallah, MEAHCO and Mouwasat) should be beneficiaries of increased private sector funding, although increased investment would be needed to meet growth in demand longer-term.

Improve public facilities, work towards privatisation

The government wants to step back from financing and providing healthcare and adopt an oversight and regulatory role. Firstly the government wants to improve the quality of care offered in the public sector, with the eventual aim of working towards privatisation of public assets and healthcare provision. This implies no privatisations short-term.

Near-term: Management contracts for public hospitals

In the near-term the government could seek to improve healthcare provision in the public system by using private sector expertise, through, e.g. contracting management of public facilities to the private sector. Individual such contracts are unlikely to transform company earnings, and the small size of existing Saudi hospital groups limits the depth of management available to capture large contracts, in our view.

Long-term: insurance coverage up to 31m from 10.5m

In the long-term a widespread adoption of private health insurance is likely to raise volumes, spurring an increase in private healthcare capacity, either from organic investment or participation in privatisations. In Saudi currently, 10.5m/31m people have insurance. We assume a more universal scheme would offer lower pricing than that today, and listed incumbents would need to make a cultural shift towards addressing this market. Such a market could attract new competitors, domestic or foreign.

Type of insurance system and reimbursement uncertain

We believe reimbursement levels would need to be known before privatisations can occur so bidders can budget and estimate their return on capital. It is still uncertain whether the government would seek to introduce insurance through the incumbent private operators (Bupa Arabia, Medgulf and Tawuniya) or set up its own insurance company e.g. an Abu Dhabi-style Daman. Nor whether it would subsidise premiums or simply force the private sector to employ more Saudis.

Government needs to improve private sector relations

At present the government has not paid private hospitals for the treatment of public patients referred to them for over a year. Until the relationship improves, engaging with the private sector could be difficult.

Healthcare implications of National Transformation Plan

The Saudi government's Vision 2030 document calls for an improvement in healthcare in Saudi Arabia. The private healthcare sector is likely to be a beneficiary both in the short-term and long-term. In the short-term the National Transformation Plan (NTP) targets the private sector to be responsible for funding 35% of healthcare spend in 2020, up from 25% today. In the longer-term, the Saudi government wants to remove itself from directly providing and financing healthcare, instead focussing on public health and the regulation of the healthcare sector. Raising private sector funding thresholds to 35% appears possible if a number of measures are used, e.g. more rigorous enforcement of requirement that Saudis working in private sector hold insurance could increase penetration by 6% alone. The quality targets set for the public sector do not appear unduly onerous in theory, but execution remains key.

What the Vision 2030 announcement said on healthcare

Transfer the responsibility for health care provision to a network of public companies

Promote competition and transparency between public and private companies to enhance the standard and quality of health care services

Prepare for privatisation in the longer term.

Work towards developing private medical insurance to improve access to medical services and reduce waiting times

Table 16: National transformation Plan objectives for the Ministry of Health

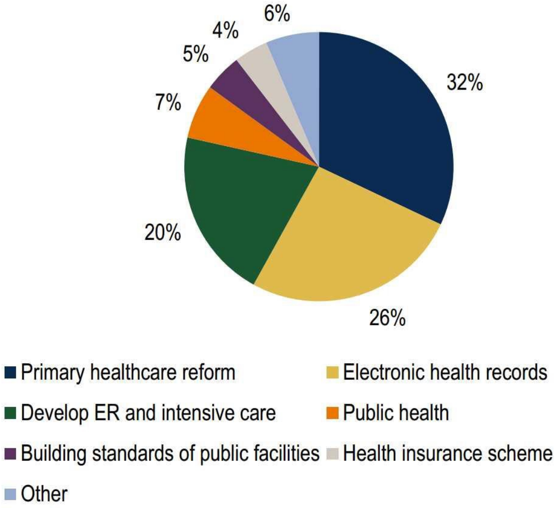
No.	Strategic objective	Key Performance Indicator (KPI)	Unit	Baseline	2020 target	Regional b'mark	Int'l b'mark
1	Increase private sector share of spending through alternative financing methods and service provision	Percentage of Private sector contribution in total healthcare spend	%	25	35	37	60
2	Increase the efficient utilization of available resources	Opex for every new inpatient admission	SAR	33,000	33,000	39,000	NA
3	Improve the efficiency and effectiveness of the healthcare sector through the use of information technology and digital transformation	Percentage of Saudi citizens who have a unified digital medical record	%	0	70	NA	100
4	Increase training and development both locally and internationally	Number of resident Saudi physicians who are enrolled in training programs	Number	2,200	4,000	NA	NA
5	Increase the attractiveness of nursing and medical support staff as a preferred career path	Number of qualified Saudis in the field of nursing and support staff for every 100,000 people	For every 100000	70	150	460	1,106
6	Improve healthcare provision before hospitalization and in the main hospitals (ER & ICU)	Percentage of patients who received emergency or urgent care with medical decision made (admission/ transfer/ discharge) in less than 4 hours in key hospitals	%	40	75	Under study	95
7	Improve integration and continuity in service provision by developing the primary care	Number of primary healthcare visits per capita	Number	2	4	3.4	7
8	Improve the infrastructure, facility management, and safety standards in healthcare facilities	Number of licensed medical facilities (affiliated with the Ministry of Health and private)	%	40	100	100	100
9	Attain acceptable waiting times across all stages of service delivery	% of appointments received in specialized medical disciplines within 4 weeks (average for all specialties in key hospitals)	%	<40	70	Under study	83
10	Improve governance in the health system in order to enhance accountability with regards to quality issues and patient safety	% of Healthcare facilities reporting comprehensive performance and quality measures	%	10	100	NA	100
11	Adopt a national plan for emergency response to public health threats per international standards	WHO emergency preparedness assessment score – average score for Riyadh, Jeddah and Eastern Province	Score	Calculation In Progress	4-5	Under study	Under study
12	Identify additional sources of revenues	Total revenue generated from private sector for utilizing government health resources	SAR bn	0.3	4	NA	NA
13	Improve public health services with focus on obesity and smoking	Increase in percentage of smoking incidence	%	Calculation In Progress	Reduce by2% from baseline	12.5*	10.5*
		Increase in percentage of obesity incidence	%	Calculation In Progress	Reduce by1% from baseline	19.4*	5*
14	Improve the quality of life and healthcare service provided to patients outside hospitals	The percentage of patients who get health care after critical care and longterm hospitalization within 4 weeks	%	25	50	NA	65
15	Improve quality and safety principles as well as skills of service providers	Percentage of hospitals that meet the US median for patient safety culture	%	10	50	NA	50

Source: Saudi National Transformation Plan

SAR23bn budget allocated for the MoH's NTP programmes

The Ministry of Health (MoH) has been allocated SAR23bn for the transformation plan over the next five years, with any contribution from the private sector on top. Over half (56%) of the funds have been allocated to reform of primary healthcare and to the establishment of electronic medical records. Only 4% of funds (SAR937m) have been allocated to health insurance.

Chart 55: The MoH has been allocated SAR23bn over five years under NTP



Source: BofA Merrill Lynch Global Research, NTP

Government currently payor and provider

The government currently directly funds public hospitals, which provide free healthcare for citizens. The typical flaw with a unified system is that increased efficiency is not necessarily sought or rewarded. Employer-provided healthcare insurance is mandatory for both Saudis and expats working in the private sector, as well as their dependents.

Significant scope for increase in insurance coverage from 10.5m to 31m people

Saudi currently has a population of 31m, of which 10.5m have health insurance, a 34% penetration rate. Bupa Arabia believes another 2.5m Saudis who work in the private sector (including dependents) don't have health insurance but should have. Assuming the government increases enforcement, near-term penetration could increase to 42% of the population even with extra reforms.

Strain in relations between government/private sector

The government has not paid private providers who have treated government patients for a year, which is hardly going to encourage the private sector's further involvement, particularly those companies that have seen the biggest increase in their receivables (Al Hammadi and Dallah). However, the government can still put considerable pressure to bear on the sector to achieve its aims if it so wishes.

Increased Saudisation could be used as a policy

The easiest way for the government to increase private financing of healthcare to 35% from 25% would be to raise Saudisation requirements. Under Saudisation, the government sets a minimum proportion of a company's staff that must be Saudi citizens, which differs by industry sector. Increased Saudi employment in the private sector short-term would put the onus of financing healthcare onto employers. However, such a move would likely pressure private sector margins.

Increase pressure on employers to trade down on insurance

The impact of Saudisation as a mechanism would likely be negative for pricing although positive for volumes, assuming the hospital has the capacity to accept more patients. Employers would likely seek to trade down in terms of the insurance schemes they offer in a bid to reduce their involuntarily raised expenditure. Insurers would seek better deals with hospitals as a result and could seek restricted network arrangements to curtail costs. Hospitals would be under pressure to be more flexible on prices to avoid any pressure should lower-cost packages impact their volumes.

Vision implies a long process

To us, the government's vision implies a two stage process over a long time-frame:

- **In the short-term:** improve the quality of care offered in the public sector, potentially with the assistance of private healthcare
- **In the long-term:** Only when public facilities have improved, privatise such facilities (although one medical city is to be privatised under a public-private partnership)
- **Concurrently:** Prepare for a roll-out of private health insurance to finance private provision

Abu Dhabi and Dubai are probably not parallels

In Abu Dhabi and Dubai, private healthcare was encouraged to avoid expats relying on public facilities. The large expat population in proportion to the locals (85:15) supported the creation of private facilities that it was then possible for the governments to fully fund Emiratis to use. The quality of public facilities was not a cause for concern - many Emiratis still prefer to use government facilities for more serious problems.

Expat population and quality of public system differ in Saudi

The proportion of expats (33% of the population) is not large enough to drive the establishment of a large, high quality private hospital base sufficient to serve the entire Saudi population. Additionally, the quality of both care and infrastructure in the public system is seen as lacking, suggesting additional investment in facilities is required.

Degree to which private sector will benefit is uncertain

Given the lack of detail provided to date, it is hard to assess the benefits, or risks to the incumbent private hospital operators in detail. We see two main ways the private sector can participate in healthcare reform:

- Win contracts to run and improve public health facilities
- Benefit from volume growth any roll-out of private health insurance could spur
- Participate in privatisations

In the long-term, volumes up but pricing down, likely to benefit nonetheless

At present there is limited supply of quality private hospitals and the incumbents are in a strong bargaining position with insurers. Greater private hospital supply, and competition, would likely pressure pricing, and margins in the longer-term, albeit offset by increased volumes. Assuming only basic services are offered to most citizens under any government-driven scheme, there would still be a place for offering higher-quality accommodation for example, to higher-income patients.

Private sector could win public hospital contracts

Private operators could be allocated management contracts for public hospitals to improve efficiency of existing hospitals, train public sector administrators and raise clinical standards. Presumably reimbursement under such contracts would be based on a fixed fee, potentially with performance clauses for improved outcomes or lowering costs. More substantive contracts, where the private operator is responsible for all aspects of operating the hospital could be more accretive, but bring greater financial risk.

Table 17: Examples of management contracts

Company	Contracts
Dallah	Dallah has had two types of hospital contracts. One, which just involves management expertise, was worth SAR4.5m over five years and 10% of the hospitals income. The other was worth SAR89m over five years but in this Dallah was responsible for all staffing and supplies and infrastructure operation and maintenance.
Mouwasat	Mouwasat managed Najd consulting hospital in Riyadh in return for management fees of c.SAR3m per year; this was subsequently changed to receiving a share of revenues and net profits.
NMC	NMC was awarded a contract to manage a government hospital in Umm Al Quwain, UAE. The five year contract sees it earn $5m (SAR18m) in fees per year, subject to certain performance criteria being met. During the term of the contract NMC is to train government staff to which it will hand over administration at expiration.

Source: BofA Merrill Lynch Global Research, company report

Private groups may lack capacity to undertake such contracts

The private hospital groups in Saudi Arabia are small. Habib Medical Group, Mouwasat and Middle Eastern Healthcare group are the largest, with only 8, 5 and 4 hospitals respectively. We would question whether any company has the depth of management to undertake management of a large number of government hospitals. Potentially the government could seek to import foreign expertise.

Insurance roll-out likely to boost private volumes

Longer-term, the roll-out of private insurance would likely have a positive impact on volumes available for incumbent operators to capture. This will depend though on the available capacity of private operators at that time and level at which reimbursement is set. The current listed operators typically target high income patients and a lot of the growth in volumes would likely be in the low-to-middle income spectrum.

Cultural shift would be needed to target larger population at lower price point

It would require a clear cultural shift for the incumbent listed hospital operators to switch from offering a high quality service to high income individuals to operating facilities that served the wider population at what we presume would be a much lower price point.

Investment is difficult until pricing and reimbursement is clarified

Until reimbursement levels are clarified, existing private operators/investors will not be able to make decisions on investing in private healthcare facilities.

Privatisations seem unlikely in the near-term

Given the stated intention to improve the quality of hospitals prior to privatisations, it seems unlikely there will be any near-term, although the NTP does call for privatisation of one of the medical cities in a public-private partnership. Saudi operates a number of medical cities, which are typically collections of hospitals that act as tertiary referral centres. The largest of these is the King Fahad medical city in Riyadh, with 1,095 beds. It is uncertain which city is planned for privatisation. As with further investment by private operators in new facilities, the level of reimbursement would need to be clarified so potential bidders can estimate their return on investment from participating.

270 public hospitals in Saudi currently

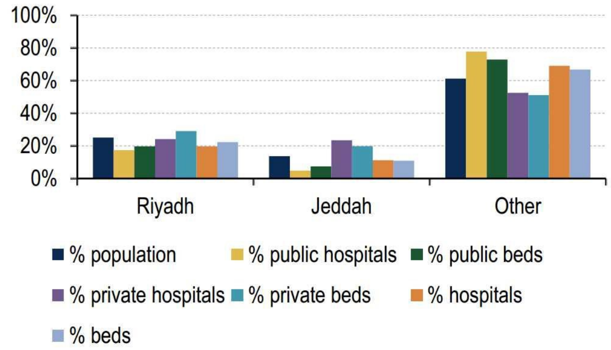
Latest available data (2014) from the Saudi Ministry of Health discloses 270 public hospitals with 40,300 beds. Of these, 47 hospitals are in Riyadh, with 25% of the population, and 13 hospitals are in Jeddah, which holds 14% of the population. Presumably these assets would be the most valuable during a privatisation process, not only because of the population sizes, but also because the provision of beds hasn't kept up with population growth, ensuring high demand. Both areas also host the highest number of private hospitals in the country.

Table 18: Public and private hospital and bed distribution

	Riyadh	Jeddah	Other
Population	7,717,467	4,224,568	18,828,340
Public			
Hospitals	47	13	210
Beds	7,937	2,993	29,370
Beds/hospital	169	230	140
Private			
Hospitals	34	33	74
Beds	4,554	3,109	8,001
Beds/hospital	134	94	108
Total			
Hospitals	81	46	284
Beds	12,491	6,102	37,371
Beds/hospital	154	133	132

Source: BofA Merrill Lynch Global Research, Ministry of Health

Chart 56: Public and private hospital and bed distribution



Source: BofA Merrill Lynch Global Research, Ministry of Health

Hospital quality would be key to attracting private buyers

The government has not invested in hospital infrastructure and some hospitals may require substantial investment by any buyer, if they are acquired at all. One Saudi hospital operator has said that there is no way it would seek to acquire government hospitals for that reason.

Bidders would likely want to obtain scale benefits from multiple purchases

Operating hospitals has benefits of scale in terms of centralising certain non-medical services (e.g. purchasing, catering, laundering) and allocating central costs over a greater revenue base. The greatest financial benefit would come from acquiring groups of hospitals to maximise these benefits. There would likely be cases where staff, and/or high net worth individuals chose to bid for individual assets.

Privatisation proceeds could defray some of NTP cost

Amounts raised from any privatisation would reflect the degree of competition for individual assets as well as level of reimbursement and the profits any acquirer could generate. In terms of pure asset values we assume the Riyadh and Jeddah hospitals (60) would be worth more than those elsewhere (210), simply because of the greater population currently and higher expected future growth. Government hospitals are likely lower cost than that of private facilities given they are not competing on the quality of accommodation offered. Arbitrarily assuming US$100m for hospitals in Jeddah and Riyadh and US$50m elsewhere would imply proceeds of $16.5bn (SAR61bn). If every hospital were sold then it could more than cover the SAR23bn allocated to the MoH for the NTP. That seems unlikely however; there will be hospitals that require substantial investment and aren't worth the indicative figures we have used above, or are in areas where profitability means there will be no return for the private sector at that level of acquisition.

Reforms could attract additional competition

Incumbent private hospital operators are best placed to expand and are seen as the natural buyers of any privatised assets. However, any substantial privatisation process, or the potential created by widespread adoption of private health insurance in general, could attract interest from foreign investors or operators. We note only a small number of international hospital groups to date have experience of successfully operating in multiple countries (IHH, Mediclinic), and we think Saudi Arabia could be too challenging as a first step for those yet to operate outside their domestic market.

Structure of insurance market needs to be set

There are a number of ways the government could roll-out private health insurance more widely:

- Set up its own insurance company

- Work with existing private insurers
- Use a back-door approach and just force private employers to hire more Saudis

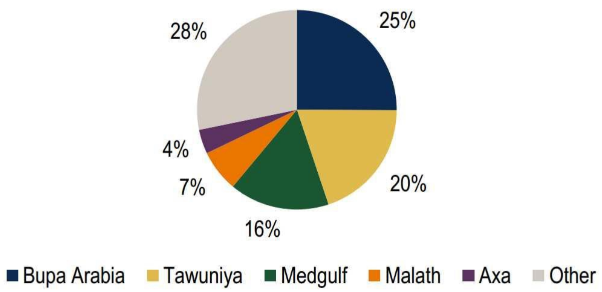
The government could set up a state insurance company

The government could set up a single-state insurer that is either funded through direct contributions or through general taxation. The government would likely need to get external help from a party knowledgeable in insurance and risk, which could be an insurer, Saudi or otherwise. Daman, Abu Dhabi's health insurer is owned 80% by its government and 20% by German insurance company Munich Re.

Existing private insurers may not have capacity if they were made responsible

The government could decide to execute any financing scheme through existing private insurers, or give citizens the option of choosing either a government or private insurer. The existing private healthcare insurers in Saudi Arabia are unlikely to have the resources (both people and systems) in place to provide cover for an additional 20m people immediately and will need some time to prepare. Any large scale immediate expansion could put them at risk of substantial underwriting losses that they would presumably seek to have back-stopped by the government in the initial stages.

Chart 57: Health insurers by share of claims (1H15)



Source: BofA Merrill Lynch Global Research, CCHI

Consumer: a necessary pain

Abdelrali El Jattari >>
Merrill Lynch (DIFC)
abdelrali.eljattari@baml.com

The Saudi government initiatives taken in the National Transformation Plan (NTP) will transform the Saudi consumer landscape. We note a mixed effect on the sector:

- On the positive side, we anticipate 1) job creation for Saudi nationals in the private sector with a big focus on Small and Medium Enterprises (SMEs); 2) a strong emphasis on education; 3) an ongoing increase of the participation of women in the workforce; 4) an increase of the percentage of self-sufficiency in broiler production which should ultimately support local producers; and, 5) a significant pick-up of the religious tourism during Haj and Umrah.

- On the negative side, we anticipate 1) a rationalization of subsidies for water and electricity; and, 2) a more competitive retail environment, more open to international players, attracting foreign direct investment (FDI).

Prefer staples over discretionary

While slowing consumer credit and weaker private sector consumption growth amid a young population do not support Saudi spending in both staple and discretionary items, we expect staple-related stocks such as Al Othaim and Savola to be better positioned to capture the marginal consumption. This is principally due to their exposure to necessities (food items) in a very fragmented market. Having said that, we expect more pressure on the opex of all Saudi consumer corporates, while revenue growth outlook will remain subdued in the short-term given the weaker consumer confidence and disposable income.

A more moderate growth outlook; Saudi female workforce drives private sector

Following a 12% 10-year CAGR fuelled by the rise in household income, we conclude that the retail sector will continue to play an important role in the rising participation of the Saudi workforce, women in particular (10,000 in 2010 vs 120,000 in 2014 according to the Ministry of Labor) supporting their disposable income in the long-term. Overall, the Saudi retail sector is one of the largest employers in Saudi Arabia with 1.5mn workers (17% of the Saudi workforce) of which low-cost foreign workers represent 80%. This implies that Saudis working in the retail account for 300,000 (40% are female), doubling in 4 years.

Needed reforms: better macro at the expense of micro

While we acknowledge that the targets set by the government are paving the way for the right reforms (development of the private sectors, a more competitive economic landscape, attraction of FDI), we expect the Saudi retail outlook to be marked by a slower growth in the coming years. In particular, we highlight few major challenges for the effective implementation of the plan: 1) incentives to enrol Saudis into consumer-related jobs suggesting rising opex pressure for consumer stocks; and, 2) 20% public payroll cut will put pressure on Saudi disposable income in the short-term given than c80% of the Saudi workforce is in the public sector.

Key strategic objectives and Key Performance Indicators (KPIs) to watch

The National Transformation Plan (NTP) outlays several strategic objectives and related KPIs to watch by ministry which we believe will impact the Saudi consumer universe:

- The Ministry of Economy and Planning is looking to 1) expand privatization of governmental services; 2) increase the efficiency of government subsidy programs; 3) establish specific zones with competitive advantages to enhance investments;

and, 4) develop tourism and entertainment sector. To do so, the ministry mainly targets to decrease the subsidy for water and electricity by SAR200bn by 2020 and to expand the private sector contribution to GDP from 40.5%.

- The Ministry of Commerce and Investment is looking to 1) reinforce the protection of the Saudi consumers; and, 2) develop SMEs by incentivizing the culture of entrepreneurship and supporting productive families. The ministry targets by 2020 to 1) more than double the number of established entities from 50k to 104k; and, 2) increase the contribution of SME to non-oil GDP from 33% to 35%.

- The Ministry of Environment, Water and Agriculture is looking to 1) optimize the use of renewable water resources for agricultural purposes; and, 2) support national companies. The ministry targets by 2020 1) to consume more than twice less water in the agricultural sector, which should be reflected in a lower percentage of water used in the agricultural sector relative to the total available renewable water resources (191% vs 416% today); and, 2) to reinforce the Saudi independence to the broiler production by increasing the percentage of self-sufficiency in broiler production to 60% from 42%.

- The Ministry of Haj and Umrah is looking to provide the opportunity for the largest number of Muslims possible to perform Haj and Umrah. The Ministry targets by 2020 the number of Haj pilgrims to increase from 1.5mn to 2.5mn and the number of Umrah Pilgrims from abroad to increase from 6mn to 15mn.

- The Ministry of Labor and Social Development is looking to 1) improve education and enable more Saudi nationals to enter the job market; and, 2) to increase the participation of women in the workforce. The ministry targets by 2020 to 1) create 1.2mn jobs in the private sector (men and women); 2) reduce the unemployment rate to 9% from 11.6%; 3) reduce the cost of employment of Saudis compared to expatriates to 280% from 400% today, which implies a higher cost of expatriates; and, 3) to increase the proportion of female force to 28% vs 23% today.

Table 19: National Transformation Plan objectives and Key Performance Indicators (KPIs) affecting the consumer sector

Strategic objective	Key Performance Indicators (KPIs)	Unit	Baseline	2020 target	Regional b'mark	Int'l b'mark
Ministry of Economy & Planning						
Increase efficiency of government subsidy programs	Value of water and electricity subsidy decrease	SARbn	0	200	N/A	N/A
Expand privatization of governmental services	Private sector contribution to GDP	%	40.5	N/A	20/27	36.4
Ministry of Commerce and Investment						
Boost entrepreneurship	Number of established entities (LLC)	#	50000	104000	N/A	347015
Boost SMEs	Contribution of SME to non-oil GDP	%	33	35	60	N/A
Ministry of Environment, Water and Agriculture						
Optimize the use of renewable water resources for agricultural purposes	Water used in the agricultural sector relative to total renewable water resources	%	416	191	42.76	19.1
Maintain security of vital resources of the country	% of self-sufficiency in broiler production	%	42	60	80	140
Ministry of Haj and Umrah						
Accommodate larger number of Muslims to perform Haj	Number of formal pilgrims (domestic and foreign)	mn	1.5	2.5	N/A	N/A
Accommodate more Muslims to perform Umrah	Number of Umrah Pilgrims from abroad	mn	6	15	N/A	N/A
Ministry of Labor and Social Development						
Provide suitable jobs for citizens	Number of suitable private sector job opportunities for Saudis	000	0	1200	N/A	N/A
Provide suitable jobs for citizens	Saudi unemployment rate	%	11.6	9	N/A	5.8
Provide suitable jobs	Cost of employment of Saudis vs expatriates	%	400	280	N/A	N/A
Empower women	Proportion of female labor force	%	23	28	N/A	N/A

Source: Saudi National Transformation Plan

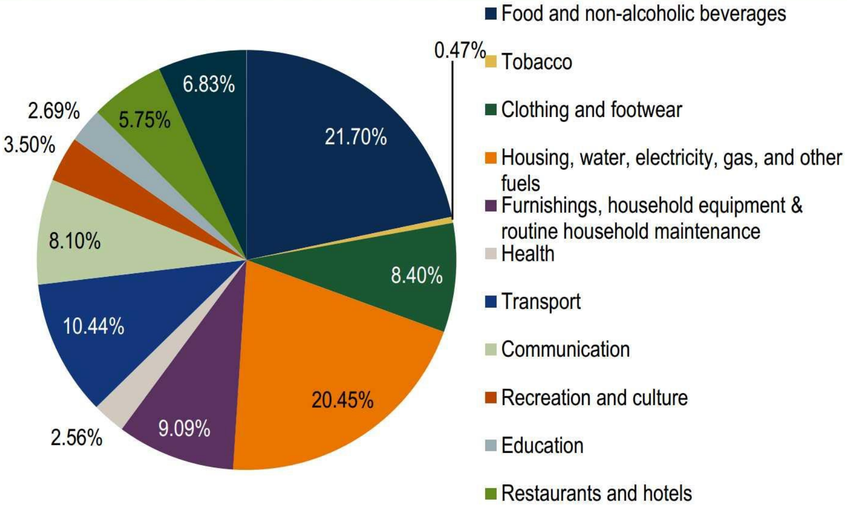
Rationalization of subsidies for water and electricity

The sweeping energy, water and electricity administered price changes introduced in December 2015 are a first step in the five-year fiscal consolidation and economic transformation strategy. We expect further reviews of energy, water, and electricity prices over the medium-term. Furthermore, the likely government review of current levels of fees and fines, introduction of new fees, application of a VAT and introduction of excise taxes on tobacco and soft drinks will also impact the Saudi consumer.

Subsidy cuts on energy and utilities to add up to 1.5ppt to CPI inflation only

We estimate the December natural gas price hike on petrochemical firms, domestic crude oil price hike as well as the combined gasoline and diesel price hike introduced could add US$2.2bn, US$2.0bn and US$3.8bn to central government revenues if fully passed to the budget (a combined 1.2% of GDP). We estimate the direct impact of these higher gasoline, water and electricity prices to add 1.3-1.5ppt to CPI inflation due to their low basket weights (c.1.5%, c0.4% and 1.6% respectively). We think a 5% VAT tax could add c2% of 2015 GDP in fiscal revenues over the medium-term.

Chart 58: Consumer Price Index (CPI) basket– consumer expenditure breakdown



Source: Haver, BofA Merrill Lynch Global Research

Subsidies account for 15% of Almarai's net income; 40% are related to poultry

Almarai continues to receive subsidies from the Saudi government, amounting to SAR295mn in 2015. This represents 15% of the Almarai's 2015 net income. The government pays subsidies based on price of subsidized feed importer, which means it may increase or decrease naturally. Regarding the breakdown by business (poultry, dairy), it is unfixed, depending on the price of feed and ingredients. As of 2015, subsidies for poultry accounted for 40% and dairy for 60%.

Saudi Arabia approves 100% foreign ownership rules

- Saudi Arabia's cabinet has approved rules governing foreign ownership of retail and wholesale businesses in the Kingdom. The regulations, which were first discussed last year, allow foreign investors to own 100% of retail and wholesale businesses in the country. Earlier, the ownership ceiling for foreigners was set at 75%. This initiative is aimed at attracting more regional and international brands to Saudi Arabia, creating more jobs, and boosting non-oil revenue. This has raised concerns regarding the sustainability of the business model of Saudi retailers.

- Real estate is key to entering a market: unlike its peers, Al Hokair enjoys favourable access to prime locations in well-located shopping malls owned by its parent shareholder, Al Hokair Group. Fashion retailers such as Inditex continue to compete for good-quality real estate in shopping centres and on high streets.

- Jarir and Extra are more vulnerable to the entry of international retailers such as Apple as barriers of entry are very low for consumer electronics and appliances. Furthermore, the government emphasis on developing tourism and family entertainments (Six Flags, Sea World) represents a potential threat to Jarir and Extra's business models which still play an 'entertaining' role in Saudi Arabia.

- Such initiatives would result in (1) reducing the number of stores; (2) a rapid modernization of the sector; and, (3) a significant increase in labor productivity by adopting merchandising best practices.

Rationale for liberalization

The rationale for permitting FDI in retail trading is (1) attracting investments in production and marketing; (2) improving the availability of such goods for the consumer; (3) encouraging increased sourcing of goods from Saudi Arabia; and, (4) enhancing competitiveness of Saudi enterprises through access to global designs, technologies and management practices.

Typically, global retailers follow a 100%-ownership business model, which explains their reluctance to establish their presence in Saudi Arabia because of the restrictive policy environment. This has been reflected in the little amount of FDI received in the Saudi retail sector.

However, we do not believe that such decision will provide foreign investors more ability to have control of a company as the current ownership ceiling of 75% already allows them to pass both ordinary and special resolutions.

Key opportunities of 100% foreign ownership initiative:

We perceive several opportunities emerging from such an initiative:

1) **Capital infusion**: FDI is one of the major sources of investments for a developing country like Saudi Arabia wherein it expects investments from multinational companies to improve economic activity, create jobs, share their expertise, back-end infrastructure and research and development in the host country.

2) **Boost competition and dampen inflation**: the entry of the many multinational corporations will promise intense competition between the different companies offering their brands in a particular product market. This will result in availability of many varieties, reduced prices, and convenient distribution of the marketing offers.

3) **Improvement of supply chain:** Improvement of supply chain/distribution efficiencies, coupled with capacity building and introduction of modern technology will help arrest wastages, particularly in the food supply chain.

4) **Jobs creation**: the entry of foreign companies into retailing in Saudi Arabia will not only create job opportunities but will also ensure quality in them, improving standards of living and life styles.

Al Hokair: Eased restrictions on foreign investors: perception vs reality

The liberalization of the Saudi retail industry raised concerns among Al Hokair investors regarding the potential loss of the Inditex franchise, which we estimate accounts for 8% of its total stores. We think this risk is overstated for the following reasons:

- **Relationship with parent**: Unlike its peers, Al Hokair enjoys favourable access to prime locations in well-located shopping malls owned by its parent shareholder, Al Hokair Group. The group owns and operates 13 shopping malls across Saudi Arabia (1.2m sqm of prime real estate) through its subsidiary Arabian Centres Company. However, we understand that rents are currently negotiated on commercial terms, thereby limiting the risk of a conflict of interest. We do not think the situation will change.

- **Real estate is key to entering the market**: Fashion retailers continue to compete for good-quality real estate in shopping centres and on high streets. Going forward, there is likely to be significant competition between these companies for new sites, especially in prime shopping centres, as they all require similar locations, typically prominent, wide-fronted premises. Zara, in particular, uses its shop windows to advertise its products.

- **Local manufacturing and distribution network would add execution risks and costs to international retailers such as Inditex**: While Saudi Arabia has not yet provided the details and conditions of the potential eased restrictions on foreign investors, we understand that the government is looking to attract investments, diversify its economy and improve Saudization. This means foreign retailers could be asked to set up local manufacturing and distribution networks within Saudi Arabia, which we believe could discourage retailers such as Inditex from entering the market directly.

More malls, less bargaining power with international retailers

The accelerating pace of mall developments in Saudi Arabia by regional competitors to Al Hokair's parent company suggests a rising risk of diminishing bargaining power for Al Hokair with its international brand partners such as Inditex. We expect Shumoul Holding 55% owned by Mabanee, the Kuwaiti mall operator, to develop a 400k sqm mall in Riyadh while Dubai-based Majid Al Futtaim recently announced that it is looking to develop two malls in Riyadh of 300k and 100k sqm of GLA respectively resulting in four times more stores (300) in the next five years. Such new entrants can then accommodate foreign retailers (such as Inditex), which could question the sustainability of Al Hokair's franchise model.

Jarir: exposed to NTP innovations

What if Apple enters Saudi Arabia

Should Apple enter Saudi Arabia, it would negatively impact the Apple resellers such as Jarir as the majority of the sales would be transferred to the Apple-branded stores. As such, we highlight 2 key risks to Jarir that would ultimately deteriorate earnings outlook: (1) rising competition from Apple resulting in a weaker footfall trend due to weaker electronic sales, which are key for store traffic; and, (2) intensifying competition from organised retail space (malls), reducing the market share of specialist retailers such as Jarir.

Jarir's earnings sensitive to electronics business

With c40% of revenues and c20% of gross profit derived from the electronics segment, Jarir remains vulnerable to the threat of foreign retailers entering directly Saudi Arabia.

We estimate that a 20% decline in electronic sales would result in a 5% direct decline in gross profit. However, this ignores the other negative effects of other product categories due to a weaker footfall trend.

More entertainment in KSA, less attraction for Jarir's stores

In line with the guidance of the NTP, we expect Saudi Arabia to develop the entertainment sector (theme parks, cinemas). This signals the emergence of new aspirations and social changes for Saudis. While these developments will support implementation of Saudi Vision 2030, they will pose threats to the existing business models of Saudi retailers such as Jarir and Extra that are currently substituting for the lack of entertainment opportunities in the country.

Real estate: NTP positive but not enough

The Saudi residential sector is suffering an acute housing shortage which we estimate at c1.2mn homes. Key reasons are:

- 1) **Land is expensive:** land constitutes as much as 50% of the cost of a housing unit which is more than double the typical 20-25% in more affordable areas globally.

- 2) **Affordability issue**: banks are reluctant to provide mortgages to workers from the private sector.

- 3) **Limited financing**: the Real Estate Development Fund (REDF) which used to provide interest-free loans to eligible Saudis for home purchase or construction, has been unable to cope with the demand for new loans and has a backlog of about 450,000 applicants. The housing ministry is looking to issue Islamic bonds to help fund the country's REDF at the end of 2017.

- 4) **Overrun development costs**: Long delays for construction approval and permits translate to overrun costs which create obstacles for developers to launch economically feasible projects.

A persisting shortage of affordable residential supply

The shortage of housing is not a new phenomenon – it has been ongoing for decades. The demand outlook from end-users is solid. Domestic household formation (marriages) is a large component: newly married Saudi couples need an affordable place to live. With market demand growing by 105K units pa (BofAMLe), we believe the shortage of supply will persist for the following reasons:

- **The limited scalability of residential projects** – Dar Al Arkan is the largest residential developer with only a 1% potential market share (delivery capability of 1,100 units pa).

- **The capital-intensive Saudi financing model** – without funding support from the government and a virtually non-existent off-plan sales market, Saudi developers have to secure funding to undertake larger residential projects. Dar Al Arkan is having to lever up, as land sales, which were once the primary source of funding, are no longer enough to enable it to scale up.

- **A lack of incentives for large regional developers to enter the market** – the dearth of off-plan sales has dried up the cash flows of developers and limited their ability to finance new projects and, in the worst cases, ongoing projects. Furthermore, the profitability of affordable housing projects is not compelling enough to attract developers to the Kingdom.

Fragmented Saudi residential market

Some 75% of new housing supply is from individuals and micro-home builders. This trend highlights Saudis' preference for building their own homes and illustrates the immature profile of the real estate market in Saudi Arabia. Sovereign actions in neighbouring countries, like the UAE, have boosted the sector (land granted to developers, capital flows from international investors). We see big developers becoming more active in the Kingdom but the effects on supply will not be immediate, in our view. So far, they represent only 15% of the total market.

Key strategic objectives and KPIs to watch

- The Ministry of Housing is looking to: 1) improve performance of the real estate sector and increase its contribution to the GDP; and, 2) enable citizens to obtain a suitable residence. The ministry targets by 2020 to: 1) double the real estate sector contribution to GDP from 5% to 10%; 2) reduce the average time required to approve and license new residential real estate; 3) reduce the housing unit cost from 10x to 5x of the gross individual annual income; and, 4) increase the real estate financing to non-oil GDP ratio from 8% to 15%.

- The Saudi Commission for Tourism and National Heritage is looking to increase and develop hospitality facilities and tourism services. The ministry targets by 2020 to increase the number of hotel rooms/apartments by 39% to 621.6k.

Positive government ambitions but…

The rise of real estate finance to non-oil GDP from 8% to 15% implies an incremental funding of US$32bn which could finance c212k additional units by 2020 or c53k additional units p.a. This compares to an estimated annual demand of 181k units which come on the top of the existing c1.2mn shortage of housing.

…needs to be combined with land tax and sukuk issuance initiatives

Based on the target Key Performance Indicators (KPIs) set for the Ministry of Housing, we estimate the initiative to be supportive but not sufficient to cater to the natural annual demand which is 3.5 times more elevated than the implied additional supply funded by the programme. We understand this initiative has to be combined with the land tax initiative and the potential sukuk issuance from the Ministry of Housing which could add further liquidity to the REDF to finance more units.

Land tax - gradual revenue source

Land tax among flagship measures adopted by the Saudi government

We expand more below on our understanding of the land tax initiative which was recently approved by the Cabinet to tackle the housing shortage and raise government revenues. Saudi Arabia's Cabinet announced in November 2015 that it will gradually levy a 2.5% tax on undeveloped land plots owned by persons or private entities in urban areas, after the country's Shura council approved a proposal to impose taxes on the so-called white lands. The draft law was amended to include gradual taxes in accordance with a specific timetable to force landlords to sell land, thus increasing land supply.

The regulations adopted in mid-June define undeveloped land as all vacant land dedicated to residential use or commercial residential use within the urban boundary limits. Fees will be applied to in a gradual series of stages; first, to owners of undeveloped land with areas exceeding 10,000 square metres; second, to single owners of developed land of more than 10,000 square metres in one master development plan; third, to single owners of developed land of more than 5,000 square metres in one master development plan; and, last, to single owners of land of more than 10,000 square metres in one city. The Ministry of Housing will determine the land locations where the tax will apply. The extension of the tax in later phases to developed land, from undeveloped ones at first, will broaden the tax base and minimize tax evasion.

Tax proceeds will be deposited in SAMA

The taxes and fines collected by the Ministry of Housing shall be deposited in the central bank, the Saudi Arabian Monetary Agency, and will be used to develop the country's infrastructure for new housing projects, according to the state-owned news agency.

Law could be effective this month, but gradual implementation likely

Under the new system, the housing ministry has been in charge of issuing executive procedures and rules for the new legislation within 180 days. The law will be effective 180 days after publication in the official gazette. The housing ministry is finalizing necessary amendments to the decree by end-June 2016 and aim for implementation by 17 January 2017 (including a 1-3 year grace period).

Tax on land: a lever to lower real estate prices in urban areas

Given that the government is giving landlords a grace period in order to either develop or sell land to developers, we perceive the newly approved land tax as a lever to lower real estate prices. Ultimately, this would improve household affordability, reduce social tensions and indirectly tackling energy subsidies as it frees up potential household income to spend on electricity and gasoline.

Land tax could generate up to US$11bn/yr from the major urban cities

We estimate that the government could potentially collect up to US$11bn per annum of tax related to undeveloped land in the top 5 urban Saudi cities which accommodate more than 50% of the Saudi population. While the law will not be effective until end-2016, we do not expect the tax proceeds to be significant early on.

Table 20: Land tax proceeds estimates for the 5 major urban cities

Land tax could generate up to US$11bn per annum in first phase

City	Selected urban land area (sqm)	White land value (SARbn)	Annual Land tax proceeds @ 2.5% (SARbn)
Riyadh	1,300,000,000	494	12.4
Jeddah	1,686,000,000	481	12
Mecca	850,000,000	485	12.1
Medina	293,000,000	88	2.2
Dammam	200,000,000	48	1.2
Total	**4,329,000,000**	**1,595**	**39.9**

Source: SAMA, BofA Merrill Lynch estimates for white land and proceeds

Table 21: NTP objectives and KPIs affecting the real estate sector

Strategic objective	KPI	Unit	Baseline	2020 target	Regional b'mark	Int'l b'mark
Ministry of Housing						
Enable citizens to obtain a house	Real estate sector contribution to GDP	%	5	10	13	20
Improve performance of real estate sector	Average time required to license new residential new projects	Day/ Permit	730	60	44	26
Enable citizens to obtain a house	Housing unit cost multiples of gross individual annual income	x	10	5	6.7	3
Enabling citizens to obtain suitable housing financing	Real estate financing to non-oil GDP	%	8	15	16	75

Source: Saudi National Transformation Plan

Table 22: NTP government costs related to the consumer and real estate sectors

Consumer	SAR (000)
Ministry of Economy and Planning	3,293,255
Ministry of Commerce and Industry	4,313,050
Ministry of Environment, Water and Agriculture	13,942,405
Ministry of Haj and Umrah	333,600
Ministry of Labor and Social Development	7,931,710
Total	**29,814,020**
Real Estate	**SAR (000)**
Ministry of Housing	59,166,666

Source: Saudi National Transformation Plan

Metals & Mining: ambitious growth target

Faisal AlAzmeh, CFA >>
Merrill Lynch KSA Company
faisal.alazmeh@baml.com

The NTP has focused on several areas relating to domestic drivers, i.e., housing, retail, trade and finance. However, a core part of the Saudi government's transformation strategy is to focus on expanding the Kingdom's mining potential. Saudi Maaden has historically been Saudi's mining champion, but the mining sector currently contributes only around 2% to the Kingdom's GDP. The NTP has set a high target for the sector: its contribution to GDP growing by 50% to SAR97bn by 2020 and sector employment increasing by 40% to 90k. While the first target could be achieved given the amount of capital that is currently being deployed in the sector, the latter part is questionable in our view as capital intensive projects are not the best means for employment generation.

Government focus on mining to boost jobs and growth

The focus on mining is driven by the ample and growing resources available. The Kingdom's largest mined products are phosphates, bauxite and gold. It supplies around 8% of global DAP production and has the largest integrated aluminium company in the world. Ma'aden continues to discover new gold and copper sites as well. We believe the outlook for mining is encouraging and the contribution to GDP targeted by the government could be achieved. We believe investments in this sector faces two hurdles:

1. Diversification – the base metal cycle lags that of oil but is eventually driven by the same global drivers (global growth). Furthermore, the NTP has currently set a low number for its capital commitments to the industry (US$82mn), achieving the targeted diversification and job creation at the moment seems to be highly dependent on attracting private sector capital. This would require attractive incentives schemes such as energy subsidies, tax exemptions or ease of issuing work permits for expats. However, at the moment, these have not been addressed by the Saudi authorities.

2. High capital intensity – for example, the government has spent close to US$21bn on Saudi Maaden but the company employs only around 6,000 permanent workers, of which Saudis account for nearly a third. While the build-up of this capacity does create a multiplier effect of almost 10x during the construction phase, most of the added jobs are usually low income blue-collar jobs that tend not to attract Saudis.

Domestic services sectors create more jobs than mining for dollars spent

We argue that capital spent on domestic services sectors is likely to yield more employment for dollars spent. A key example is Saptco, the Saudi Public transport Company, which has spent around US$450mn on assets and currently employs around the same number of permanent employees as Maaden, of which 1,200 are Saudis. We acknowledge that the technology and technical transfer will be different in an investment like Maaden versus Saptco, yet we believe the economic efficiency of using the mining sector as a means of generating employment is questionable.

NTP is already happening: Saudi Aramco has project to create 2,000 jobs

The plan to develop the mining industry is moving forward. Saudi Aramco has already announced its plans to build a downstream industrial complex along with GE and Cividale SpA. The complex will include the first-of-its-kind high-end forging & casting manufacturing facility and will serve the region's maritime and energy industries. The project will cost around US$400mn and likely generate around 2,000 jobs.

Exports will largely depend on market conditions

Part of the NTP relies on growing Saudi exports, including that of the mining industry. While it is easy build a plant, the timing of when there projects come online and how it coincides with the cycle matters. The ability to dump will largely depend on the position on the cost curve, which seems to be fading with the recent reduction of subsidies.

The conglomerate model to continue

Saudi Arabia has historically adapted the conglomerate model to build up its industrial sectors, such as petrochemicals (SABIC), metals & mining (Maaden) and steel (Hadeed). We believe the government continues to focus on this model in its NTP plan and will use a number of vehicles to expand the sectors in question: Maaden in metals & mining and the new military company in defence.

Maaden has transformed from a pure gold producer to one of the largest conglomerates in Saudi Arabia at the moment. Through three key Joint-Ventures (JVs), the company is a top producer of DAP (Diammonium Phosphate) and gold and aluminium. The NTP clearly sets the stage for Maaden to grow even further but is unclear on how this is to be achieved and what it would mean for shareholders. Also, in our view, the all-in-one model has delivered subpar returns globally and the trend is towards the streamlining of operations and specialised vehicles. As an example, the largest petrochemical and agrichem companies, Dow and Dupont, plan to merge and then split into three different businesses, providing more specialisation. We can see the advantages of the specialised entity model as it provides cost synergies, efficiency and a clear strategy for the company. Also, in the conglomerate model, certain parts of the company are often dwarfed as other revenues streams represent a larger contribution to returns. Board and management could be more or solely focused on the sub-segment had if it is a stand-alone entity they have to monitor and manage, in our view.

Table 23: National Transformation Plan objectives in the mining sector

	Baseline	Target 2020	Regional benchmark	Global benchmark
Overall kingdom:				
Number of jobs created	65k	90k	75k	210k
Contribution to GDP	SAR64bn	SAR97bn	SAR13bn	SAR262bn
Number of localized technologies	Under study	125	NA	NA
Less developed regions in Saudi:				
Number of jobs created	0	12k	NA	NA
Expected contribution from private sector	0	SAR28bn	NA	NA
memo:				
Initiative cost to the government: SAR310mn (US$83mn)				

Source: Saudi National Transformation Plan

Oil & gas/petchems: focus on downstream

Faisal AlAzmeh, CFA >>
Merrill Lynch KSA Company
faisal.alazmeh@baml.com

Focus on downstream expansion is a priority

An increased focus on downstream expansion is a strategic development priority in Saudi Arabia, with the aim of exploiting competitive advantages on feedstock costs and the strategic location. Furthermore, the government sees it as a mechanism to create job opportunities for nationals. We see the NTP natural gas target as optimistic as it faces many hurdles, while imports could be a better alternative which authorities now suggest they could be open to eventually.

Downstream chemicals represents the second largest focus of the NTP

The NTP is firmly focused on the integration of chemicals/refining/mining capacity into value-add downstream operations that would limit imports of certain products and provide technology transfer into the Kingdom. This accompanies the government's plans to expand its natural gas and refining capacity by 48% and 13%, respectively.

The NTP includes a sizeable allocation to the Royal Commission for Jubail and Yanbu, the government body that overlooks, promotes and develops the petrochemicals/energy-intensive/mining sectors in the Kingdom. The government plans to spend around US$10.9bn on developing various areas including: 1) the oil-to-olefins (OTC) complex in Yanbu; 2) colleges and institutes in Jubal Industrial area; 3) the necessary infrastructure in the community area of Yanbu Industrial City; and 4) value-add transformation industries in Raas AlKhair Industrial City.

Table 24: NTP focus for the Royal Commission of Jubail and Yanbu

Key focus areas	US$mn
Development of the OTC project	772
Development of colleges and institutes in Jubail Industrial City	731
Development of infrastructure in Yanbu	1,003
Development of residential areas in Jubail Industrial City	642
Development of Multi-Modal Logistics Hub in Yanbu Industrial City	543
Development, protection and rehabilitation of public facilities in Jubail Industrial City	954
Development of value-added transformation industries in Raas Alkhair Industrial City	805
Construction of Housing in Jubail Industrial City	538
Development of Mineral industries port in Yanbu Industrial City	556
Other infrastructure and projects (38 projects, size between 100mn to 500mn)	4,557
Total	**11,101**

Source: Saudi National Transformation Plan

Subsidies are likely to be needed

A key obstacle to building more refining capacity and a large OTC complex is the high capital intensity of such projects and the subsidies required to ensure an acceptable rate of return. We see the natural gas target as optimistic with many hurdles; importation could be a better alternative. We believe the government plans to reduce subsidies in water and electricity, but maintain them for feedstock prices as intensive natural gas production requires cheap gas to thrive. However, a move towards benchmarking to US gas prices is a likely and logical step, in our view. An alternative scenario would to mark natural gas prices in the Kingdom to the LNG netback out of Qatar. This would facilitate an additional US$2.5bn from the sector but would also put several companies at risk of becoming loss making (especially companies with a high content of heavy feed in their input mix). Should such a scenario take place, we believe implementation would be carried out over phases in order to allow companies to cope with the new pricing order.

Table 25: Additional costs on petrochemical producers of higher feedstock prices

	Current scheme (US$50/bbl)	US Spot/(US$50/bbl)	Government revenues raised (US$bn)
Ethane	950	1,411	461
Methane	201	457	256
Propane	3,918	4,898	980
Butane	789	986	197
Naphtha	2,348	2,934	587
Ammonia	64	132	69
Feedstock cost	8,269	10,819	2,550
Electricity & water	1,039	2,053	1,014
Total	**9,309**	**12,872**	**3,564**

Source: IHS Chemical, BofA Merrill Lynch Global Research. US spot refers to natural gas prices of US$2.6/mn BTU. The US$2.5bn feedstock price rise could directly accrue to the central government through Saudi Aramco, while higher electricity and water charges would not.

Moving further down the value chain

Downstream operations in Saudi have historically been international Joint-Ventures (JVs) and government-owned entities like SABIC. In July 1993, the government issued a Royal Decree that merged all the state-owned refineries, distribution activities and marketing operations under Saudi Aramco.

This resulted in a transfer of the government's stake in three key refineries and international refining and petrochemical operations to one state-owned company. Today, Saudi Aramco is one of the largest crude oil refiners in the world and aims to become a top-three petrochemical producer through standalone petrochemicals facilities like Sadara, downstream integration at its many refineries, and potential acquisitions. The recent announcement of the US split with Shell was followed by comments that the breakup was due to different strategies by both companies with Aramco increasing its focus on petrochemicals.

Petrochemicals - a route to diversification

Saudi Arabia is emerging as one of the largest petrochemicals producers globally. The country initially started with integrated refining leading to petrochemical production through its multiple JVs in Saudi and other regions, and is now building the largest petrochemical complex globally. One of the goals is to maximise the value of the Saudi hydrocarbon chain. Ultimately, this expansion into petrochemicals should further support the diversification drive and thus support the stated goals for the Saudi economy.

Saudi Arabia (potentially through Saudi Aramco) aims to become a leader not only in commodity chemicals, but also downstream petrochemical conversion. In the cities of the Jubail and Rabigh, the Ministry of Energy, Industry and Mineral Resources (along with Saudi Aramco) is constructing value-parks next to upstream chemical facilities. The aim is to provide integration and the infrastructure for SMEs to operate and produce value-added products in the Kingdom. This diversification into specialty chemicals could increase returns from the current US$500/ton level to around US$2000/ton by 2040, according to local press citing senior Aramco officials. Saudi Arabia could also look to grow its international refining footprint to provide further integration opportunities, as was suggested by the recent split of the Motiva JV (Saudi is more keen to expand its downstream operations, whilst Shell is looking to reduce them).

Two petrochemical projects are also being built. The first is the Sadara project in Jubail, which is expected to become the world's largest integrated chemicals complex with 3MMt of output. It will use naphtha as feedstock. The second is Petro Rabigh II, which is the expansion project of the existing Petro Rabigh plant that will process 4MMt (93kb/d) of naphtha feedstock, and is planned to be launched in 2016.

NTP includes the long discussed Oil-to-Olefins (OTC)

Saudi's national transformation plan includes the development of a US$770mn industrial cluster whereby expansion to the Aramco refinery is likely to take place, providing the platform for the long discussed OTC, Oil-to-Olefins, project. Saudi authorities have highlighted in several press releases that Aramco and SABIC are looking at potential oil-to-olefins projects. The Kingdom has a proprietary technology that enables it to maximise the yield of olefins per ton of oil used. According to SABIC's management, this project has been carried out on a small scale at test sites and it is conducting feasibility studies at a larger-scale facility. Aramco has also highlighted plans in this area. While both companies declined to comment, Bloomberg highlighted that sources indicate they are assessing the possibility of embarking on a project together. Such a step would be the first of its kind in Saudi Arabia, with Aramco and SABIC working alongside each other on such a large scale.

Refining capacity target: almost there

The NTP aims to increase the refining capacity to 3.3mn bpd per day by 2020 from 2.9mn bpd today. Saudi has recently completed and launched two large refining projects with 400kbpd installed refining capacity each: (1) SATORP in Jubail launched in June 2014; and (2) YASREF in Yanbu – first shipment made in January 2015 and inaugurated in January 2016. In addition, it is building a new refinery, Jazan, which should process 400kb/d of Arabian heavy and medium crude oil; launch is expected in 2019.

Table 26: Saudi Arabia government refining assets

Name	Location	Completion	Refining capacity, kb/d	Saudi Aramco's share in capacity, kb/d	Saudi Aramco's ownership (%)	Partner
Wholly-owned domestic refineries						
Jeddah	Jeddah	1967	90	90	100%	n/a
Yanbu	Yanbu	1979	240	240	100%	n/a
Riyadh	Riyadh	1981	126	126	100%	n/a
Ras Tanura	Ras Tanura	1986	550	550	100%	n/a
Total			**1,006**	**1,006**		
Domestic refining JVs:						
SAMREF	Yanbu	1983	400	150	50.0%	ExxonMobil
SASREF	Jubail	1986	305	114	50.0%	Shell
Petro Rabigh I	Rabigh	1990	400	150	37.5%	Sumitomo (37.5%, free float 25%)
SATORP	Jubail	2014	400	250	62.5%	Total (37.5%)
YASREF	Yanbu	2015	400	250	62.5%	Sinopec (37.5%)
Total			**1905**	**914**		
Total in Saudi Arabia			**2,911**	**1,920**		
International refining JVs:						
S-Oil	South Korea	1991	669	424	63.4%	S-Oil
Motiva	USA	2002	1070	535	50.0%	Shell
Showa Shell	Japan	2004	445	67	15.0%	Shell
Fujian	China	2007	280	70	25.0%	Sinopec, ExxonMobil
Total			**2464**	**1096**		
Total capacity globally			**5,375**	**3,016**		

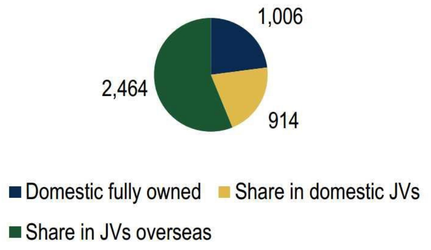
Source: Saudi Aramco, Energy Policy, BofA Merrill Lynch Global Research

Table 27: Saudi Arabia petroleum product output in 2014 (mn bpd)

	Saudi Aramco ownership (%)	LPG	Naphtha	Gasoline	Jet Fuel	Diesel	Fuel oil	Asphalt	Total
Wholly-owned domestic refineries									
Jeddah	100%	0.9	2.9	4.0	0.0	2.4	9.2	6.4	25.8
Yanbu	100%	2.4	3.2	11.5	-0.4	29.3	30.9	0.0	77.0
Riyadh	100%	1.8	0.0	10.9	2.7	19.2	0.0	6.5	41.2
Ras Tanura	100%	5.0	15.0	43.9	7.8	76.1	32.5	7.2	187.4
Subtotal		**10.1**	**21.1**	**70.3**	**10.0**	**127.1**	**72.6**	**20.1**	**331.3**
Saudi Aramco share in domestic JVs									
SAMREF	50%	-1.1	0.0	25.0	11.1	18.8	14.4	0.0	68.2
SASREF	50%	1.3	11.7	2.2	9.4	14.2	13.3	0.0	52.2
Petro Rabigh	38%	1.3	7.3	6.9	4.9	11.7	13.4	0.0	45.5
SATORP	63%	1.3	3.9	11.1	8.2	31.6	7.5	0.0	63.7
Subtotal		**2.9**	**22.9**	**45.2**	**33.7**	**76.3**	**48.6**	**0.0**	**229.5**
Total		**13.0**	**44.0**	**115.6**	**43.7**	**203.4**	**121.2**	**20.1**	**560.9**

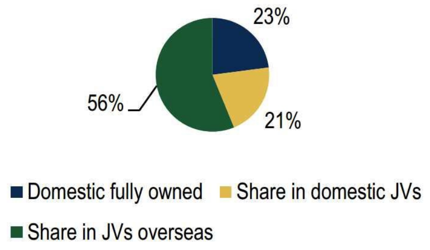
Source: Saudi Aramco, BofA Merrill Lynch Global Research

Chart 59: Saudi government refining assets (effective stake in k bpd)



Source: Saudi Aramco, BofA Merrill Lynch Global Research

Chart 60: Saudi refining breakdown per region (effective stake)

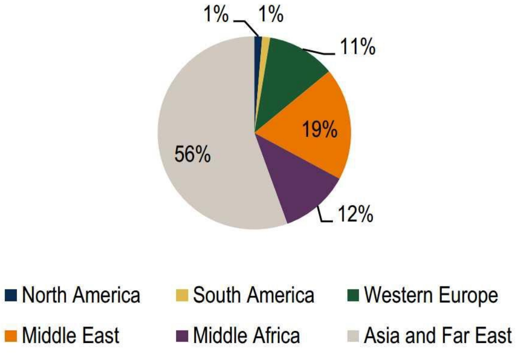


Source: Saudi Aramco, BofA Merrill Lynch Global Research

Exports are geared towards Asia and Far East

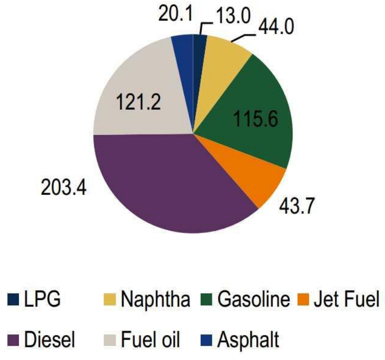
Saudi Arabia exported c.1mbpd of oil products in 2014, mainly to Asia and the Far East (56%), other Middle Eastern countries (19%) and Western Europe (12%).

Chart 61: Oil product export volumes split per region in 2014



Source: Ministry of Energy, Industry and Mineral Resources.

Chart 62: Saudi Arabia oil product slate in 2014



Source: Company data, BofA Merrill Lynch Global Research

Ambitious natural gas expansion

The NTP also focuses on expanding natural gas production. The government aims to grow the existing production by 48%, largely through non-traditional routes. According to the minister of Energy, Industry and Mineral Resources, natural gas accounts for 50% of energy production and the target is to increase this to 70% by 2020 through expanding local production or importing if necessary. The government's priority will be to explore the option of local production, although imports were, for the first time, not excluded. We consider this as an interesting development as importing LNG into Saudi costs the same as producing it internally through non-traditional routes, in our view. The Saudi statement acknowledges that Saudi Arabia is willing to be open to the possibility should imports provide a more economic means for this purpose.

Table 28: National Transformation Plan objectives in the petrochemicals, oil, and refining sectors

	Base line	Target 2020	Regional benchmark	Global benchmark
Oil production capacity (mn bbls/day)	12.5	12.5	3.8	11
Dry gas production capacity (bn scf per day)	12	17.8	5.7	16
Refining capacity	2.9	3.3	1.1	1.9
memo:				
Initiative cost to the government (total SAR491mn or US$131mn):				
Energy	SAR216mn			
Oil & Gas	SAR200mn			
Gas network	SAR75mn			

Source: Saudi National Transformation Plan

Utility sector: sharing the capex burden and achieving cost reflective tariffs

Ali Dhaloomal
MLI (UK)
ali.dhaloomal@baml.com

Eight strategic objectives aimed at being self-funding

The electricity sector is mentioned in eight strategic objectives of the NTP 2020. The two most important measures in our view are the one related to subsidies cuts and the one on the liberalisation of the sector which should free out financial resources to fund the future capex burden as well as the other targets under the NTP. This is even more relevant as the NTP doesn't provide any additional funding for the sector.

Unanswered questions as to impact on electricity sector

While we think these initiatives should be welcomed as they would reduce the financial burden on Saudi Electricity Company (SEC) from its ambitious capacity expansion plan (SAR240bn / US$64bn to be spent over the 2016-21, mostly in generation), some questions remain unanswered, especially in regards to SEC's capital structure and financial sustainability on a standalone basis. It remains unclear if the proceeds from the disposals of the minority stakes in the GenCos will be used to fund future capex plans or to alleviate SEC's commercial debt (US$16.3bn as of FYE15) and government and government-related debt and payables (c.US$46bn as of FYE16). Also, it remains unclear if on the back of the NTP, SEC would start paying its current payables (fuel sourced from Aramco, electricity sourced from SWCC and fees owed to municipalities) instead of accruing these amounts as it does now.

1. **Saving SAR200bn annually by cutting the water and electricity subsidies:** The current print of the NTP doesn't give any detail about how this target will be split between water and electricity subsidies, but we note that the plan aims to reach full cost reflective water tariffs by 2020 (from 30% cost coverage at the moment). Recall that starting from January 1st, electricity prices have been increased by a weighted average of close to 20%, with residential prices (half of the consumption) now ranging SAR 0.05-0.30 / KWh vs SAR 0.05-0.26 previously. The highest increases are for the commercial sector, the governmental entities and the agricultural sector. SEC's management recently indicated to us that they expect a neutral impact over the long term from this decision as it mirrors an increase in feedstock price charged by Aramco. However, over the short to medium term, SEC cash flows are expected to benefit from the tariff decision given that the company books all its fuel costs in payables without disbursing any amount to Aramco.

2. **Increasing the percentage of power generation through strategic partners to 100% from 27% now:** We believe that the aim here is to bring industrial or financial partners in all SEC's and Saline Water Conversion Corporation (SWCC) generation units, in line with the electricity sector reform which expected to be implemented by late 2016 / early 2017. According to SEC's management, the most likely scenario involves SEC's generation assets to be split into four separate regional GenCos where minority stakes would be sold to major global utilities or sold in the public market.

3. **Increasing the efficient use of "fuel" in power generation to 40% from 33% now:** We think that this objective entails increasing the share of efficient combined cycle gas turbines (CCGT) plants. Given that SEC currently represents 73% of the country's installed capacity and that 19% of its electricity is generated via efficient combined cycle gas turbines (CCGT), this means that 71% of the remaining tier party capacity is "efficient". We note that another 39% of SEC's installed generation capacity is made of less efficient open cycle gas turbines (OCGT).

4. **Increasing the reserve to peak demand to 12% from 10% now:** This reserve margin was just at 2.3% back in 2012 and SEC was assigned a target of 10% which management expects to fulfil. In light of the recent improvements and SEC's ambitious capacity expansion plan (SAR240bn / (US$64bn in capex over the 2016-21), the target looks realistic to us.

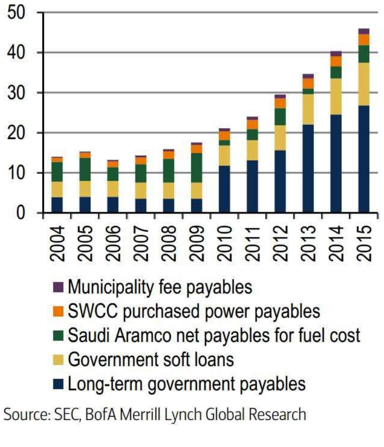
5. **Reducing the outages of more than 5min to 3.0 per year from 6.4 per year now and the average outage time to 120 minutes from 262 minutes now.** This goal will likely be achieved through the ongoing investments in expanding and improving SEC's transmission and distribution network (half of the overall capex incurred in 2015 went to these two segments).

6. **Increasing the access to electricity to 99.5% of the population from 99% now:** It is hard to quantify the cost of this objective, but we believe that it entails increasing the reach to the remote Eastern and Southern regions of the country where the economics of expanding the transmission network are not relevant.

7. **Reach 86 GW of installed capacity vs. 69 GW at the end of 2015 (incl. 50 GW at SEC alone):** The target is in line with SEC own current ambitious expansion plan (22.8GW to be added over five years to reach 68.3 GW by the end of 2018) and factors in additional capacity from SWCC and IPPs as well.

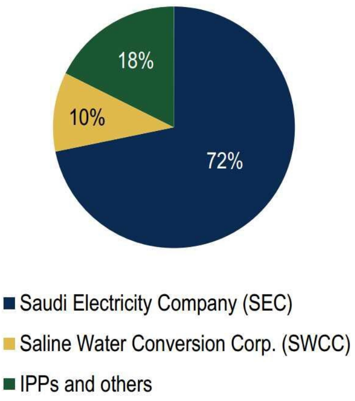
8. **Generate 4% of the electricity from renewables (equiv. to 3.45 GW):** we would expect the projects to be primarily based on onshore wind and solar. It is likely that these projects will be fully developed by IPPs under long term power purchase agreements where SEC would be the offtaker and assume the higher cost of the electricity from these ReIPP. While SEC has currently no renewable generation capacity at all, it has recently invited expressions of interest for potential lead developers for two 50MW solar photovoltaic projects, which we see as a (modest) step towards the NTP target. Based on the KACARE study on grid impact, Saudi Arabia's electricity network should be able to integrate up to 14GW of renewables capacity by 2020. Finally, we note that ACWA Power, the largest privately-owned utility company in Saudi Arabia and one of the main IPP players in the country (along with Engie), has developed a proven expertise in wind and solar projects internationally. ACWA is likely to be one of the main players in the upcoming liberalisation plans.

Chart 63: SEC's debt and payables to the gov. & GREs grew at 17.4% CAGR since 2009 (US$bn)



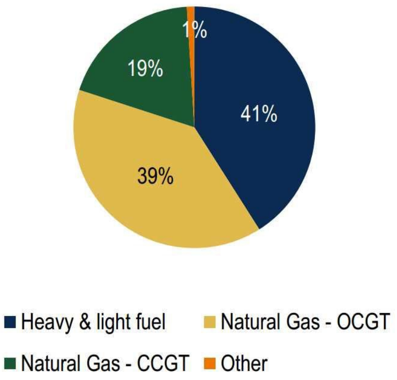
Source: SEC, BofA Merrill Lynch Global Research

Chart 64: Saudi Arabia power generation installed capacity split in 2015 (total: 69 GW)



Source: SEC, BofA Merrill Lynch Global Research

Chart 65: Saudi Electricity Company (SEC) fuel mix for electricity generation



Source: SEC

Defence: Vision 2030 supports defence spending

Celine Fornaro >>
MLI (UK)
celine.fornaro@baml.com

Benjamin Heelan >>
MLI (UK)
benjamin.heelan@baml.com

Saudi Arabia is increasing its regional role as a Security and Defence nation in the Middle East region. The 2030 Vision includes a higher local content in defence procurement which can be achieved through new contract awards and increased cooperation with large Original Equipment Manufacturers (OEMs). BAE is best positioned, in our view: it has been the in the country since 1966 with 5,000 local employees and Saudi Arabia represents 21% of BAE's group sales.

Saudi Arabia to build out military industrial base

Defence is one of the key industries that Saudi is pinpointing for investment as part of its future strategy plans, and one where it is well placed to become a global leader, in our view. Saudi Arabia has one of the largest defence budgets globally, and we expect its focus on defence investment to remain broadly intact regardless of the price of oil. Due to regional tensions, defence spending will likely remain high for the foreseeable future, in our view. Saudi Vision 2030 and the associated National Transformation Plan (NTP) will likely increase focus on use of defence to achieve their economic diversification goals. This is likely to take place through capability deployment and employment in partnership with long term defence partners. Saudi authorities aim to raise the locally-sourced defence procurement from 2% today to 50% in 2030.

Much of the countries' defence spending is for foreign imports. According to consultancy IHS, 1 out of every 7 dollars spent on defence imports in 2015 was spent by Saudi Arabia, which was predicted by IHS to increase its defence imports by 52%yoy to US$9.8bn in 2015.

In our view, Saudi strategy to increase locally manufactured defence content is justified, given their level of purchasing power with international contractors. However, this could trigger a round of new contracts' awards to set up new terms. The State owned Taqnia company has been signing some contracts with international companies but this remains very limited at this stage.

Saudi commitment to defence spending positive for European Defence

Saudi Arabia's Ministry of Finance 2016 defence budget spending remains high at SAR213bn (cUS$57bn) which is 25% of the total budget. In 2015, the budgeted defence spending reached SAR307bn (US$82bn). Despite the large decrease in headline defence and security expenditure, budgetary figures suggest the allocation for the Ministry of Interior, Ministry of National Guard, Ministry of Defence, General Intelligence Directorate, and Saudi Royal Guard Regiment could be heavily supported from elsewhere in the budget, if required. Figures released by the Ministry of Finance suggest that SAR183.0bn of the state budget for 2016 - c21% of total budgeted spending - will be allocated through a new 'Budget Support Provision' that is to be used "to give more flexibility to the budget and redirect capital and operational expenditure in line with national priorities and to fulfil spending requirements". In our view, the commitment to defence spending is positive for companies who have presence in the area, BAE (Buy, 600p) & Thales (Buy, €87).

BAE Systems ties to Saudi Arabia go as far back as 1966, when the company was contracted to Lightning and Strikemaster aircraft and equipment in a programme called the 'Magic Carpet'. In 1985, a new government-to-government (UK/Saudi Arabia) agreement was signed covering supply and support of Tornado, Hawk and PC-9 aircraft. A new government-to-government agreement, known as the Saudi British Defence Co-operation Programme, was inaugurated at the start of 2007. Salam Project, which provides for the acquisition of 72 Typhoon aircraft in a programme aimed at the modernisation of the Saudi Armed Forces, was launched in September 2007. This carries substantial packages of work and expertise transfers, aimed at developing the Saudi industrial defence sector. BAE Systems has made a major commitment to the training of local nationals to bring them into management, technical and other qualified positions. Saudis constitute 21% of BAE's group sales, 57% of the total BAE Systems' workforce in Saudi Arabia making the Company one of the largest private sector employers of Saudis with 5,000 employees based locally. The UK and Saudi have discussed in the past moving the final assembly of some of the 72 Typhoons in Saudi Arabia but it failed. Now, the two countries may be trying to have locally produced Hawks on the back of the 22 trainer order at the end of 2015, according to local press.

Thales (10% of group sales in the Middle East, mainly in defence) has an established presence in the Kingdom of Saudi Arabia, with 750 employees and a diversified portfolio of activities. The company has built a mature relationship with Saudi Arabia based on solid, country-wide partnerships. Thales's major commercial successes include the Shahine and Crotale air defence contracts with logistic support. In naval defence, Thales is the prime contractor for the Sawari II contract (three frigates), that follows on from the successful Sawari I programme (four frigates) for which Thales now provides logistic support. In 2007, Thales was selected by the Saudi Arabian Ministry of Defence to supply a major electronic protection system to the Royal Saudi Air Force.

US contractors such as Lockheed Martin (supplier of F16 jets, C130J transport planes), General Dynamics (military vehicles) and Raytheon (missiles) have some local presence in Saudi Arabia, but BAE is by in large the largest and historical local presence.

Table 29: Stocks mentioned

Name	Symbol	Opinion	Q-R-Q	Price	PO
Al Hammadi	XBQYF	Buy	C-1-7	SAR 40.19	SAR 64.0
Al Othaim	XWPJF	Buy	C-1-7	SAR 99.99	SAR 121.0
BAE SYSTEMS	BAESF	Buy	A-1-7	GBP 483.4	GBP 600
BAE SYSTEMS	BAESY	Buy	A-1-7	USD 26.17	GBP 34.26
SABIC	XAUBF	Buy	C-1-8	SAR 80.87	SAR 99.5
Saudi Telecom	XUTUF	Buy	C-1-7	SAR 63.95	SAR 81.0
Savola	XSAVF	Buy	C-1-7	SAR 35.95	SAR 46.0
THALES	THLEF	Buy	C-1-7	EUR 73.5	EUR 87.0
YANSAB	XUYNF	Buy	C-1-7	SAR 40.28	SAR 48.0
Zain KSA	XOCTF	Buy	C-1-9	SAR 7.83	SAR 12.2
Almarai	XALRF	Neutral	B-2-7	SAR 53.7	SAR 61.0
Jarir	XJRIF	Neutral	C-2-7	SAR 118.64	SAR 134.0
SAFCO	XDUAF	Neutral	C-2-8	SAR 59.73	SAR 72.0
Dallah Healthcare	XJEFF	Underperform	C-3-7	SAR 86.12	SAR 79,0
Dar Al Arkan	XARKF	Underperform	C-3-9	SAR 6.49	SAR 7.0
Extra	XYDUF	Underperform	C-3-7	SAR 26.72	SAR 24.0

Source: BofA Merrill Lynch Global Research

Price objective basis & risk

Al Hammadi (XBQYF)

Our SAR64 PO for Al Hammadi is based on our DCF-derived valuation. This is equivalent to 32x FY17E EPS. Strong earnings growth means out-year valuations appear much more reasonable than near-term valuations (e.g. implied 13x FY20E EPS). Company specific risks are: lower-than-expected growth resulting from delays in executing on expansion plans, inability to raise prices to offset cost inflation, failure to manage growth without impacting margins, pricing pressure from insurers, new competition, sustained delays to government payments, and any potentially dilutive acquisitions. General/macro risks are: sustained low oil prices and any resultant slow-down in the economy, civil unrest, ongoing failure of the government to pay bills.

Al Othaim (XWPJF)

We use a combination of DCF, DDM and peer group analysis to value Al Othaim. We have taken a weighted average of values implied by our DCF, DDM and peer group analyses. We attribute a 50% weight to the DCF and 25% weights for both the DDM and the P/E multiple, which points to a PO of SAR121. We attribute a greater weight to the DCF valuation as we believe food retailers offer good visibility on cash flow generation. In the case of Al Othaim, we think DCF better reflects the stock's growth opportunity. We assume a perpetuity growth rate of 2% and WACC of 9.1% in our DCF.

The downside risks to our investment case are:
- Weaker-than-expected margin performance
- Cannibalisation effect in Saudi Arabia
- Competition: We expect competition to intensify among the large organized players, such as Savola-owned Panda which have aggressive retail expansion plans growth plan
- Saudization: Most of Al Othaim's labors are low level staff working in the warehouses and branches which depends heavily on expatriate labors. This suggests the Saudization of the staff could put further pressure on margins.

Almarai (XALRF)

We arrive at our SAR61/share price objective using a combination of P/E multiples and DCF valuation, taking the average yielded by the two methods. More specifically:

On a multiples basis, we value Almarai on 21x 2017E EPS, a 15% premium to Global (blue chip) food producer averages. We believe that at this level it is fairly valued given it is trading at a premium to its historical trading average and the weaker consumer outlook.

Our DCF analysis yields a valuation of SAR59/share, based on a WACC of 8% and terminal growth rate of 3.5%.

Risks to our PO are company specific factors such as plant outages, disease outbreaks in its biological assets, delays in the investment program and factors affecting suppliers. Macro level risks are commodity prices, government regulation, subsidy removals and price pressures from growing competition.

BAE SYSTEMS (BAESF / BAESY)

Our PO of 600p (US$34.26/ADR) is an average based on DCF, 2016-17E P/E and EV/EBITA and Sum-of-Parts. We regard as BAE's peers Lockheed Martin, LLL, General Dynamics, Northrop Grumman, Raytheon, QinetiQ, Cobham, Ultra Electronics.

Our DCF valuation (8.7% WACC and 1.8% long-term growth rate in line with defence peers and 10.7% long term margin) implies a 550p share price. Our SOP suggests at valuation range of 592-595p, using peer group EV/EBITA multiples of 12.7-13.7x and EV/sales of 1.2x. Applying EU and US peers multiples on EV/EBITA and P/E yields a valuation range of 671-679p.
At 600p BAE would be trading on 14.6x 2016E P/E and at 13.9x 2016E EV/EBITA. This P/E is in the middle of the historical range of c.20x to c.6x, and the EV/EBITA is near the high end of the historical range of c.12x to c.4x.

Upside risks are: 1) An export contract for Typhoons, in particular an order for 60+ airplanes from Saudi, 2) USD strength against the UK, 3) development of a large scale and long term conflict.
Downside risks are: 1) significant decrease to the US or UK defence budgets, below our current assumptions, 2) significant weakness of the US$ vs Sterling.

Dallah Healthcare (XJEFF)

We set a SAR79 price objective for Dallah, which is equivalent to 23x FY17E EPS. We expect earnings to double between FY15A and FY20E as beds triple. The multiple therefore falls quickly and on FY20E earnings, when some of this growth is likely to have come through, the stock trades in-line with other hospital stocks, both EM and Saudi.

Upside risks are: greater price increases than assumed in forecasts, faster uptake of new capacity, accretive M&A activity.

Down side risks are: Slowdown in the economy, lower-than-expected growth resulting from delays in executing on expansion plans, inability to raise prices to offset cost inflation, failure to manage growth without impacting margins

Dar Al Arkan (XARKF)

Our PO of SAR7.0/share is based on sum-of-the-parts methodology. We employ a WACC of 10.7%, derived from a COE and COD of 11.1% and 9.0%, respectively. Our SOTP includes (1) the NPV of the current development portfolio to 2016E, (2) the NPV of recurring cash flows from investment portfolio + the NPV of its terminal value minus incremental capex to complete the investment programme, (3) the NPV of land sales from the developed land bank, and (4) the book value (valued at cost) of DAAR s remaining land bank.
Downside risks: (1) Deterioration of demand for land would increase the perceived liquidity risk (2) Deterioration of credit environment (3) Emergence

of sub-developers which would inflate the acq. cost of land and consequently depreciate the residual value of land due to lower margins (4) Mismanagement of cash flows: a mismanagement of the resources (overspending in CAPEX for land for example) could arise liquidity concerns because of over-exposure to land sales mkt. Upside risks: A lower-than-expected borrowing cost: We forecast an incremental cost of debt of 9%, A improvement in the credit environment sensitive to exogenous factors such as liquidity and the risk appetite of the international mkt. would lower DAAR s cost of funding, We see upside valuation risk in the value of the land bank should the company manages to develop housing units with the support of the Ministry of Housing. debt-financed acquisition to boost its recurring income.

Extra (XYDUF)

We derive a PO of SAR 24 using a DCF valuation model which we believe best captures differing capital costs and growth profiles across the MENA region. Key assumptions are an 11% WACC and a 2% perpetuity growth rate. Upside/downside risks to our PO are better/worse returns from better/worse like-for-like sales and shorter/longer break even times from new international markets.

Jarir (XJRIF)

We derive a PO of SAR134 using a DCF valuation model which we believe best captures the company's plan to add c.60% new stores by 2017. Key assumptions are: - a 5-year CAGR in sales of 11% followed by an five-year CAGR of 6% and a perpetuity growth rate of 2%, - an average EBIT margin of 13%, - a WACC of 9.5% with a beta of 0.9x. Our WACC is calculated using a RFR of 5.0% and an ERP of 6.0%. We used a 2% terminal growth rate.

The risks to our PO are company-specific issues such as a failure to deliver the expected 11% top-line growth or a faster-than-expected deterioration in electronics margins. In addition, there are risks associated with a slowdown in the economy or consumer spending.

Saudi Arabian Fertilizer Company (XDUAF)

We apply a justified P/E multiple to derive SAFCO's PO of SAR72. The P/E is based on a normalized RoE of 30.2%, Cost of Equity of 10.4% and payout of 93.5%.

Upside risks to our price objective are: (1) delays in global nitrogen fertilizers capacity expansions, which would result in a tighter supply of urea and effectively higher prices, (2) Stronger demand for fertilizers and effectively prices, (3) An increase in the marginal producer cost that effectively leads to a higher urea price floor.

Downside risks are: (1) lower prices of urea due to weaker than expected demand, (2) Delay in SAFCO 5 expansion project, (3) an increase in natural gas cost.

Saudi Basic Industries Corporation (XAUBF)

We apply a justified P/E multiple to derive SABIC's PO of SAR99.5. The P/E is based on a normalised RoE of 15.8%, Cost of Equity of 11.5% and payout of 65.0%

Downside risks to our price objective are a decline in the supply of low-cost feedstock, a lower-than-expected recovery in petrochemicals prices, delays in the ramp-up of newly established subsidiaries, or weakness in steel demand in Saudi Arabia.

Saudi Telecom Company (STC) (XUTUF)

We derive our SAR81/share PO for STC on a sum of the parts basis, using a combination of DCF and market valuations for its core subsidiaries and associates, adjusting for ownership stakes. Specifically, we use DCF to value its core operations in Saudi Arabia (9.5% WACC), Viva Kuwait (10.5% WACC), and other subsidiaries (10.1% WACC). We assume a terminal growth rate of 2% across all of the markets, starting from 2023. For STC's associates, we value its holdings in Oger Telecom and Binariang using a combination of our Research team's valuation (for Oger's stake in Turk Telecom) and book value. We then add STC's net cash position at YE 2015 in deriving our PO.

Risks to our PO come from potential market share loss as both Mobily and Zain KSA, plus several MVNO's are aiming to take market share from STC domestically, following recent cut in MTR's. Additionally, we see ongoing risk from FX exposure at their subsidiary investments, especially their indirect stake in Turk Telekom.

Savola (XSAVF)

Given the diversified nature of the Savola group, we use a sum of the parts valuation in deriving our SAR46/share price objective for Savola. Specifically, we:

Value the food business on 12x 2016 earnings, a 30% discount to global peers on account have having slightly slower growth and risks presented by having a significant position in Iran (which accounts for 13% of revenues).

Value retail(Panda) at 15x 2016 earnings, a c.10% discount to global peers on account of having slower near term growth. we believe near term growth will be impacted by continued aggressive expansion and slowing demand trends in the Saudi market (following the introduction of the 2016 budget).

The investments business is valued using a variety of methods. For its stake in Almarai, we value it in line with our price objective for the shares. Savola's interests in other listed entities (Herfy, KEC and EEC) are valued at current market valuation given we do not have research coverage on the names. The stakes in none listed entities including the real estate book and private equity holdings are valued at 2x book value (YE2015).

Downside risks to our PO include higher cost escalation than we forecast, further adverse moves in commodity prices , adverse FX movements in countries outside of Saudi Arabia where Savola does business and any disruptions to its material Iranian business.

THALES (THLEF)

To derive our Thales price objective of €87, we use a 2017E sum of the parts valuation, using Rockwell Collins, Raytheon, Northrop Grumman, Ultra, and the EU Civil Aerospace peer group as peers for the Aerospace/defence segments, and Ansaldo STAS as a peer for the transport business.

At €87 Thales would trade on 18.1x 2017E P/E and 12.3x EV/EBITA which we think is appropriate given balance sheet/ portfolio optionality and building order momentum.

Upside/Downside risks to our price objective are: 1) changes in the company's free float, 2) assets swap and portfolio changes could add more value than we currently estimate 3) lower/more French defence cuts than we currently estimate.

YANSAB (XUYNF)

We apply a justified P/E multiple to derive YANSAB's PO of SAR48. The P/E is based on a normalized RoE of 14.5%, Cost of Equity of 11% and payout of 70%

Downside risks to our price objective are: 1. Lower-than-expected demand for polyester would adversely impact MEG prices and effectively earnings 2. Unexpected change in feedstock prices could negatively impact earnings 3. Unexpected shutdowns that would result in lower production.

Zain KSA (XOCTF)

We use a DCF valuation model to derive our price objective of SAR 12.2. Key assumptions are a discount rate of 11.0%, underpinned by a risk free rate of 5.5%, a cost of equity of 29% and a 6.5% cost of debt. We explicitly forecast free cash flows until 2022 and then assume perpetuity growth of 3%, as we do when valuing other Gulf operators. Our forecasts take in to account the repayment of the deferred royalty loan afforded to Zain in 2013 (repayments starting from 2021).

Downside risks are greater pricing pressure from MVNOs, the loss of its appeal against a SAR620mn tax claim, an unfavourable ruling against Zain KSA in its arbitration process with Mobily and an inability to grow market share and margins, which in turn would threaten capital requirements.

Analyst Certification

We, Abdelrali El Jattari, Ali Dhaloomal, Anton Fedotov, Benjamin Heelan, Celine Fornaro, Faisal AlAzmeh, CFA, Francisco Blanch, Hootan Yazhari, CFA, Jamie Clark, CFA and Jean-Michel Saliba, hereby certify that the views each of us has expressed in this research report accurately reflect each of our respective personal views about the subject securities and issuers. We also certify that no part of our respective compensation was, is, or will be, directly or indirectly, related to the specific recommendations or view expressed in this research report.

Special Disclosures

Some of the securities discussed herein should only be considered for inclusion in accounts qualified for high risk investment.

Disclosures

Important Disclosures

Credit opinion history

Bahrain / BHRAIN

Sovereign	Date^	Action	Recommendation
Bahrain / BHRAIN	12-Nov-2015	Initial	Marketweight

Table reflects credit opinion history as of previous business day's close. ^First date of recommendation within last 36 months. The investment opinion system is contained at the end of the report under the heading "BofA Merrill Lynch Credit Opinion Key."

Dubai / DUGB

Sovereign	Date^	Action	Recommendation
Dubai / DUGB	12-Nov-2015	Initial	Marketweight
	08-Jan-2016	Downgrade	Underweight
	28-Apr-2016	Upgrade	Marketweight

Table reflects credit opinion history as of previous business day's close. ^First date of recommendation within last 36 months. The investment opinion system is contained at the end of the report under the heading "BofA Merrill Lynch Credit Opinion Key."

Qatar / QATAR

Sovereign	Date^	Action	Recommendation
Qatar / QATAR	12-Nov-2015	Initial	Marketweight
	08-Jan-2016	Downgrade	Underweight
	26-Feb-2016	Upgrade	Marketweight
	25-May-2016	Restricted	NA

Table reflects credit opinion history as of previous business day's close. ^First date of recommendation within last 36 months. The investment opinion system is contained at the end of the report under the heading "BofA Merrill Lynch Credit Opinion Key."

BofA Merrill Lynch Credit Opinion Key

BofA Merrill Lynch Global Research provides recommendations on an issuer's bonds (including corporate and sovereign external debt securities), capital securities, equity preferreds and CDS as described below. Convertible securities are not rated. An issuer level recommendation may also be provided for an issuer as explained below. BofA Merrill Lynch Global Research credit recommendations are assigned using a three-month time horizon.

Issuer Recommendations: If an issuer credit recommendation is provided, it is applicable to bonds and capital securities of the issuer except bonds and capital securities specifically referenced in the report with a different credit recommendation. Where there is no issuer credit recommendation, only individual bonds and capital securities with specific recommendations are covered. CDS and equity preferreds are rated separately and issuer recommendations do not apply to them.

BofA Merrill Lynch Global Research credit recommendations are assigned using a three-month time horizon:

Overweight: Spreads and /or excess returns are likely to outperform the relevant and comparable market over the next three months.

Marketweight: Spreads and/or excess returns are likely to perform in-line with the relevant and comparable market over the next three months.

Underweight: Spreads and/or excess returns are likely to underperform the relevant and comparable market over the next three months.

BofA Merrill Lynch Global Research uses the following rating system with respect to **Credit Default Swaps (CDS):**

Buy Protection: Buy CDS, therefore going short credit risk.

Neutral: No purchase or sale of CDS is recommended.

Sell Protection: Sell CDS, therefore going long credit risk.

Sovereign Investment Rating Distribution: Global Group (as of 31 Mar 2016)

Coverage Universe	Count	Percent	Inv. Banking Relationships*	Count	Percent
Buy	5	12.82%	Buy	2	40.00%
Hold	28	71.79%	Hold	12	42.86%
Sell	6	15.38%	Sell	6	100.00%

* Issuers that were investment banking clients of BofA Merrill Lynch or one of its affiliates within the past 12 months. For purposes of this Investment Rating Distribution, the coverage universe includes only Sovereign issuer recommendations. A Sovereign issuer rated Overweight is included as a Buy, a Sovereign issuer rated Marketweight is included as a Hold, and a Sovereign issuer rated Underweight is included as a Sell.

FUNDAMENTAL EQUITY OPINION KEY: Opinions include a Volatility Risk Rating, an Investment Rating and an Income Rating. *VOLATILITY RISK RATINGS*, indicators of potential price fluctuation, are: A - Low, B - Medium and C - High. *INVESTMENT RATINGS* reflect the analyst's assessment of a stock's: (i) absolute total return potential and (ii) attractiveness for investment relative to other stocks within its *Coverage Cluster* (defined below). There are three investment ratings: 1 - Buy stocks are expected to have a total return of at least 10% and are the most attractive stocks in the coverage cluster; 2 - Neutral stocks are expected to remain flat or increase in value and are less attractive than Buy rated stocks and 3 - Underperform stocks are the least attractive stocks in a coverage cluster. Analysts assign investment ratings considering, among other things, the 0-12 month total return expectation for a stock and the firm's guidelines for ratings dispersions (shown in the table below). The current price objective for a stock should be referenced to better understand the total return expectation at any given time. The price objective reflects the analyst's view of the potential price appreciation (depreciation).

Investment rating	Total return expectation (within 12-month period of date of initial rating)	Ratings dispersion guidelines for coverage cluster*
Buy	≥ 10%	≤ 70%
Neutral	≥ 0%	≤ 30%
Underperform	N/A	≥ 20%

* Ratings dispersions may vary from time to time where BofA Merrill Lynch Research believes it better reflects the investment prospects of stocks in a Coverage Cluster.

***INCOME RATINGS*, indicators of potential cash dividends, are: 7 - same/higher (dividend considered to be secure), 8 - same/lower (dividend not considered to be secure) and 9 - pays no cash dividend.** *Coverage Cluster* is comprised of stocks covered by a single analyst or two or more analysts sharing a common industry, sector, region or other classification(s). A stock's coverage cluster is included in the most recent BofA Merrill Lynch report referencing the stock.

Price charts for the securities referenced in this research report are available at http://pricecharts.baml.com, or call 1-800-MERRILL to have them mailed.

Credit Opinion History Tables for the securities referenced in this research report are available at http://pricecharts.baml.com, or call 1-800-MERRILL to have them mailed.

One or more analysts responsible for covering the securities in this report owns stock of the covered issuer: Genel Energy.

MLPF&S or an affiliate was a manager of a public offering of securities of this issuer within the last 12 months: Qatar.

The issuer is or was, within the last 12 months, an investment banking client of MLPF&S and/or one or more of its affiliates: BAE SYSTEMS, Dubai, Qatar, Qatar (Central Bank), SABIC, Saudi Telecom, THALES, Yansab.

MLPF&S or an affiliate has received compensation from the issuer for non-investment banking services or products within the past 12 months: BAE SYSTEMS, Bahrain, Bahrain (Treasury), Dubai, Qatar, Qatar (Central Bank), SABIC, Yansab.

The issuer is or was, within the last 12 months, a non-securities business client of MLPF&S and/or one or more of its affiliates: BAE SYSTEMS, Bahrain, Bahrain (Treasury), Dubai, Qatar, Qatar (Central Bank), SABIC, Yansab.

In the US, retail sales and/or distribution of this report may be made only in states where these securities are exempt from registration or have been qualified for sale: Al Hammadi, Al Othaim, Almarai, BAE SYSTEMS, Dallah Healthcare, Dar Al Arkan, Extra, Jarir, SABIC, SAFCO, Saudi Telecom, Savola, THALES, Yansab, Zain KSA.

MLPF&S or an affiliate has received compensation for investment banking services from this issuer within the past 12 months: BAE SYSTEMS, Dubai, Qatar, Qatar (Central Bank), SABIC, THALES, Yansab.

MLPF&S or an affiliate expects to receive or intends to seek compensation for investment banking services from this issuer or an affiliate of the issuer within the next three months: BAE SYSTEMS, Dubai, Qatar, Qatar (Central Bank), SABIC, Saudi Telecom, THALES, Yansab.

MLPF&S or one of its affiliates has a significant financial interest in the fixed income instruments of the issuer. If this report was issued on or after the 15th day of the month, it reflects a significant financial interest on the last day of the previous month. Reports issued before the 15th day of the month reflect a significant financial interest at the end of the second month preceding the report: BAE SYSTEMS, Qatar.

MLPF&S or one of its affiliates is willing to sell to, or buy from, clients the common equity of the issuer on a principal basis: BAE SYSTEMS.

MLPF&S or one of its affiliates trades or may trade as principal in the debt securities (or in related derivatives) that are the subject of this research report: Bahrain, Bahrain (Treasury), Dubai, Qatar.

The issuer is or was, within the last 12 months, a securities business client (non-investment banking) of MLPF&S and/or one or more of its affiliates: BAE SYSTEMS, Bahrain, Bahrain (Treasury), Dubai, Qatar, Qatar (Central Bank), SABIC, THALES, Yansab.

BofA Merrill Lynch Research Personnel (including the analyst(s) responsible for this report) receive compensation based upon, among other factors, the overall profitability of Bank of America Corporation, including profits derived from investment banking. The analyst(s) responsible for this report may also receive compensation based upon, among other factors, the overall profitability of the Bank's sales and trading businesses relating to the class of securities or financial instruments for which such analyst is responsible.

BofA Merrill Lynch Global Credit Research analysts regularly interact with sales and trading desk personnel in connection with their research, including to ascertain pricing and liquidity in the fixed income markets.

Other Important Disclosures

Prices are indicative and for information purposes only. Except as otherwise stated in the report, for the purpose of any recommendation in relation to: (i) an equity security, the price referenced is the publicly traded price of the security as of close of business on the day prior to the date of the report or, if the report is published during intraday trading, the price referenced is indicative of the traded price as of the date and time of the report; or (ii) a debt security (including equity preferred and CDS), prices are indicative as of the date and time of the report and are from various sources including Bank of America Merrill Lynch trading desks.

The date and time of completion of the production of any recommendation in this report shall be the date and time of dissemination of this report as recorded electronically on the platform in which the report was accessed.

This report may refer to fixed income securities that may not be offered or sold in one or more states or jurisdictions. Readers of this report are advised that any discussion, recommendation or other mention of such securities is not a solicitation or offer to transact in such securities. Investors should contact their BofA Merrill Lynch representative or Merrill Lynch Financial Global Wealth Management financial advisor for information relating to fixed income securities

Rule 144A securities may be offered or sold only to persons in the U.S. who are Qualified Institutional Buyers within the meaning of Rule 144A under the Securities Act of 1933, as amended.

SECURITIES DISCUSSED HEREIN MAY BE RATED BELOW INVESTMENT GRADE AND SHOULD THEREFORE ONLY BE CONSIDERED FOR INCLUSION IN ACCOUNTS QUALIFIED FOR SPECULATIVE INVESTMENT.

Recipients who are not institutional investors or market professionals should seek the advice of their independent financial advisor before considering information in this report in connection with any investment decision, or for a necessary explanation of its contents.

The securities discussed in this report may be traded over-the-counter. Retail sales and/or distribution of this report may be made only in states where these securities are exempt from registration or have been qualified for sale.

This report, and the securities discussed herein, may not be eligible for distribution or sale in all countries or to certain categories of investors.

Individuals identified as economists do not function as research analysts under U.S. law and reports prepared by them are not research reports under applicable U.S. rules and regulations.

Macroeconomic analysis is considered investment research for purposes of distribution in the U.K. under the rules of the Financial Services Authority.

BofA Merrill Lynch Global Research policies relating to conflicts of interest are described at http://go.bofa.com/coi.

"BofA Merrill Lynch" includes Merrill Lynch, Pierce, Fenner & Smith Incorporated ("MLPF&S") and its affiliates. Investors should contact their BofA Merrill Lynch representative or Merrill Lynch Global Wealth Management financial advisor if they have questions concerning this report. "BofA Merrill Lynch" and "Merrill Lynch" are each global brands for BofA Merrill Lynch Global Research.

Information relating to Non-US affiliates of BofA Merrill Lynch and Distribution of Affiliate Research Reports:

MLPF&S distributes, or may in the future distribute, research reports of the following non-US affiliates in the US (short name: legal name, regulator): Merrill Lynch (South Africa): Merrill Lynch South Africa (Pty) Ltd., regulated by The Financial Service Board; MLI (UK): Merrill Lynch International, regulated by the Financial Conduct Authority (FCA) and the Prudential Regulation Authority (PRA); Merrill Lynch (Australia): Merrill Lynch Equities (Australia) Limited, regulated by the Australian Securities and Investments Commission; Merrill Lynch (Hong Kong): Merrill Lynch (Asia Pacific) Limited, regulated by the Hong Kong Securities and Futures Commission (HKSFC) and the Hong Kong Monetary Authority (HKMA); Merrill Lynch (Singapore): Merrill Lynch (Singapore) Pte Ltd, regulated by the Monetary Authority of Singapore (MAS); Merrill Lynch (Canada): Merrill Lynch Canada Inc, regulated by the Investment Industry Regulatory Organization of Canada; Merrill Lynch (Mexico): Merrill Lynch Mexico, SA de CV, Casa de Bolsa, regulated by the Comisión Nacional Bancaria y de Valores; Merrill Lynch (Argentina): Merrill Lynch Argentina SA, regulated by Comisión Nacional de Valores; Merrill Lynch (Japan): Merrill Lynch Japan Securities Co., Ltd., regulated by the Financial Services Agency; Merrill Lynch (Seoul): Merrill Lynch International Incorporated (Seoul Branch) regulated by the Financial Supervisory Service; Merrill Lynch (Taiwan): Merrill Lynch Securities (Taiwan) Ltd., regulated by the Securities and Futures Bureau; DSP Merrill Lynch (India): DSP Merrill Lynch Limited, regulated by the Securities and Exchange Board of India; PT Merrill Lynch (Indonesia): PT Merrill Lynch Indonesia, regulated by Otoritas Jasa Keuangan (OJK); Merrill Lynch (Israel): Merrill Lynch Israel Limited, regulated by Israel Securities Authority; Merrill Lynch (Russia): OOO Merrill Lynch Securities, Moscow, regulated by the Central Bank of the Russian Federation; Merrill Lynch (Turkey I.B.): Merrill Lynch Yatirim Bank A.S., regulated by the Capital Markets Board of Turkey; Merrill Lynch (Turkey Broker): Merrill Lynch Menkul Değerler A.Ş., regulated by the Capital Markets Board of Turkey; Merrill Lynch (DIFC): Merrill Lynch International (DIFC Branch), regulated by the Dubai Financial Services Authority (DFSA); Merrill Lynch (Spain): Merrill Lynch Capital Markets Espana, S.A.S.V., regulated by Comisión Nacional del Mercado De Valores; Merrill Lynch (Brazil): Bank of America Merrill Lynch Banco Multiplo S.A., regulated by Comissão de Valores Mobiliários; Merrill Lynch KSA Company, Merrill Lynch Kingdom of Saudi Arabia Company, regulated by the Capital Market Authority.

This information: has been approved for publication and is distributed in the United Kingdom (UK) to professional clients and eligible counterparties (as each is defined in the rules of the FCA and the PRA) by MLI (UK) and Bank of America Merrill Lynch International Limited, which are authorized by the PRA and regulated by the FCA and the PRA, and is distributed in the UK to retail clients (as defined in the rules of the FCA and the PRA) by Merrill Lynch International Bank Limited, London Branch, which is authorized by the Central Bank of Ireland and subject to limited regulation by the FCA and PRA - details about the extent of our regulation by the FCA and PRA are available from us on request; has been considered and distributed in Japan by Merrill Lynch (Japan), a registered securities dealer under the Financial Instruments and Exchange Act in Japan; is distributed in Hong Kong by Merrill Lynch (Hong Kong), which is regulated by HKSFC and HKMA; is issued and distributed in Taiwan by Merrill Lynch (Taiwan); is issued and distributed in India by DSP Merrill Lynch (India); and is issued and distributed in Singapore to institutional investors and/or accredited investors (each as defined under the Financial Advisers Regulations) by Merrill Lynch International Bank Limited (Merchant Bank) (MLIBLMB) and Merrill Lynch (Singapore) (Company Registration Nos F 06872E and 198602883D respectively). MLIBLMB and Merrill Lynch (Singapore) are regulated by MAS. Bank of America N.A., Australian Branch (ARBN 064 874 531), AFS License 412901 (BANA Australia) and Merrill Lynch Equities (Australia) Limited (ABN 65 006 276 795), AFS License 235132 (MLEA) distribute this report in Australia only to 'Wholesale' clients as defined by s.761G of the Corporations Act 2001. With the exception of BANA Australia, neither MLEA nor any of its affiliates involved in preparing this research report is an Authorised Deposit-Taking Institution under the Banking Act 1959 nor regulated by the Australian Prudential Regulation Authority. No approval is required for publication or distribution of this report in Brazil and its local distribution is by Merrill Lynch (Brazil) in accordance with applicable regulations. Merrill Lynch (DIFC) is authorized and regulated by the DFSA. Research reports prepared and issued by Merrill Lynch (DIFC) are done so in accordance with the requirements of the DFSA conduct of business rules. Bank of America Merrill Lynch International Limited, Frankfurt Branch (BAMLI Frankfurt) distributes this report in Germany and is regulated by BaFin.

This research report has been prepared and issued by MLPF&S and/or one or more of its non-US affiliates. MLPF&S is the distributor of this research report in the US and accepts full responsibility for research reports of its non-US affiliates distributed to MLPF&S clients in the US. Any US person receiving this research report and wishing to effect any transaction in any security discussed in the report should do so through MLPF&S and not such foreign affiliates. Hong Kong recipients of this research report should contact Merrill Lynch (Asia Pacific) Limited in respect of any matters relating to dealing in securities or provision of specific advice on securities. Singapore recipients of this research report should contact Merrill Lynch International Bank Limited (Merchant Bank) and/or Merrill Lynch (Singapore) Pte Ltd in respect of any matters arising from, or in connection with, this research report.

General Investment Related Disclosures:

Taiwan Readers: Neither the information nor any opinion expressed herein constitutes an offer or a solicitation of an offer to transact in any securities or other financial instrument. No part of this report may be used or reproduced or quoted in any manner whatsoever in Taiwan by the press or any other person without the express written consent of BofA Merrill Lynch.

This research report provides general information only. Neither the information nor any opinion expressed constitutes an offer or an invitation to make an offer, to buy or sell any securities or other financial instrument or any derivative related to such securities or instruments (e.g., options, futures, warrants, and contracts for differences). This report is not intended to provide personal investment advice and it does not take into account the specific investment objectives, financial situation and the particular needs of any specific person. Investors should seek financial advice regarding the appropriateness of investing in financial instruments and implementing investment strategies discussed or recommended in this report and should understand that statements regarding future prospects may not be realized. Any decision to purchase or subscribe for securities in any offering must be based solely on existing public information on such security or the information in the prospectus or other offering document issued in connection with such offering, and not on this report.

Securities and other financial instruments discussed in this report, or recommended, offered or sold by Merrill Lynch, are not insured by the Federal Deposit Insurance Corporation and are not deposits or other obligations of any insured depository institution (including, Bank of America, N.A.). Investments in general and, derivatives, in particular, involve numerous risks, including, among others, market risk, counterparty default risk and liquidity risk. No security, financial instrument or derivative is suitable for all investors. In some cases, securities and other financial instruments may be difficult to value or sell and reliable information about the value or risks related to the security or financial instrument may be difficult to obtain. Investors should note that income from such securities and other financial instruments, if any, may fluctuate and that price or value of such securities and instruments may rise or fall and, in some cases, investors may lose their entire principal investment. Past performance is not necessarily a guide to future performance. Levels and basis for taxation may change.

Futures and options are not appropriate for all investors. Such financial instruments may expire worthless. Before investing in futures or options, clients must receive the appropriate risk disclosure documents. Investment strategies explained in this report may not be appropriate at all times. Costs of such strategies do not include commission or margin expenses.

BofA Merrill Lynch is aware that the implementation of the ideas expressed in this report may depend upon an investor's ability to "short" securities or other financial instruments and that such action may be limited by regulations prohibiting or restricting "shortselling" in many jurisdictions. Investors are urged to seek advice regarding the applicability of such regulations prior to executing any short idea contained in this report.

This report may contain a trading idea or recommendation which highlights a specific identified near-term catalyst or event impacting a security, issuer, industry sector or the market generally that presents a transaction opportunity, but does not have any impact on the analyst's particular “Overweight” or “Underweight” rating (which is based on a three month trade horizon). Trading ideas and recommendations may differ directionally from the analyst's rating on a security or issuer because they reflect the impact of a near-term catalyst or event.

Certain investment strategies and financial instruments discussed herein may only be appropriate for consideration in accounts qualified for high risk investment.

Foreign currency rates of exchange may adversely affect the value, price or income of any security or financial instrument mentioned in this report. Investors in such securities and instruments effectively assume currency risk.

UK Readers: The protections provided by the U.K. regulatory regime, including the Financial Services Scheme, do not apply in general to business coordinated by BofA Merrill Lynch entities located outside of the United Kingdom. BofA Merrill Lynch Global Research policies relating to conflicts of interest are described at http://go.bofa.com/coi.

MLPF&S or one of its affiliates is a regular issuer of traded financial instruments linked to securities that may have been recommended in this report. MLPF&S or one of its affiliates may, at any time, hold a trading position (long or short) in the securities and financial instruments discussed in this report.

BofA Merrill Lynch, through business units other than BofA Merrill Lynch Global Research, may have issued and may in the future issue trading ideas or recommendations that are inconsistent with, and reach different conclusions from, the information presented in this report. Such ideas or recommendations reflect the different time frames, assumptions, views and analytical methods of the persons who prepared them, and BofA Merrill Lynch is under no obligation to ensure that such other trading ideas or recommendations are brought to the attention of any recipient of this report.

In the event that the recipient received this report pursuant to a contract between the recipient and MLPF&S for the provision of research services for a separate fee, and in connection therewith MLPF&S may be deemed to be acting as an investment adviser, such status relates, if at all, solely to the person with whom MLPF&S has contracted directly and does not extend beyond the delivery of this report (unless otherwise agreed specifically in writing by MLPF&S). MLPF&S is and continues to act solely as a broker-dealer in connection with the execution of any transactions, including transactions in any securities mentioned in this report.

Copyright and General Information regarding Research Reports:

Copyright 2016 Bank of America Corporation. All rights reserved. This research report is prepared for the use of BofA Merrill Lynch clients and may not be redistributed, retransmitted or disclosed, in whole or in part, or in any form or manner, without the express written consent of BofA Merrill Lynch. BofA Merrill Lynch research reports are distributed simultaneously to internal and client websites and other portals by BofA Merrill Lynch and are not publicly-available materials. Any unauthorized use or disclosure is prohibited. Receipt and review of this research report constitutes your agreement not to redistribute, retransmit, or disclose to others the contents, opinions, conclusion, or information contained in this report (including any investment recommendations, estimates or price targets) without first obtaining expressed permission from an authorized officer of BofA Merrill Lynch.

Materials prepared by BofA Merrill Lynch Global Research personnel are based on public information. Facts and views presented in this material have not been reviewed by, and may not reflect information known to, professionals in other business areas of BofA Merrill Lynch, including investment banking personnel. BofA Merrill Lynch has established information barriers between BofA Merrill Lynch Global Research and certain business groups. As a result, BofA Merrill Lynch does not disclose certain client relationships with, or compensation received from, such issuers in research reports. To the extent this report discusses any legal proceeding or issues, it has not been prepared as nor is it intended to express any legal conclusion, opinion or advice. Investors should consult their own legal advisers as to issues of law relating to the subject matter of this report. BofA Merrill Lynch Global Research personnel's knowledge of legal proceedings in which any BofA Merrill Lynch entity and/or its directors, officers and employees may be plaintiffs, defendants, co-defendants or co-plaintiffs with or involving issuers mentioned in this report is based on public information. Facts and views presented in this material that relate to any such proceedings have not been reviewed by, discussed with, and may not reflect information known to, professionals in other business areas of BofA Merrill Lynch in connection with the legal proceedings or matters relevant to such proceedings.

Any information relating to the tax status of financial instruments discussed herein is not intended to provide tax advice or to be used by anyone to provide tax advice. Investors are urged to seek tax advice based on their particular circumstances from an independent tax professional.

The information herein (other than disclosure information relating to BofA Merrill Lynch and its affiliates) was obtained from various sources and we do not guarantee its accuracy. This report may contain links to third-party websites. BofA Merrill Lynch is not responsible for the content of any third-party website or any linked content contained in a third-party website. Content contained on such third-party websites is not part of this report and is not incorporated by reference into this report. The inclusion of a link in this report does not imply any endorsement by or any affiliation with BofA Merrill Lynch. Access to any third-party website is at your own risk, and you should always review the terms and privacy policies at third-party websites before submitting any personal information to them. BofA Merrill Lynch is not responsible for such terms and privacy policies and expressly disclaims any liability for them.

All opinions, projections and estimates constitute the judgment of the author as of the date of the report and are subject to change without notice. Prices also are subject to change without notice. BofA Merrill Lynch is under no obligation to update this report and BofA Merrill Lynch's ability to publish research on the subject issuer(s) in the future is subject to applicable quiet periods. You should therefore assume that BofA Merrill Lynch will not update any fact, circumstance or opinion contained in this report.

Certain outstanding reports may contain discussions and/or investment opinions relating to securities, financial instruments and/or issuers that are no longer current. Always refer to the most recent research report relating to an issuer prior to making an investment decision.

In some cases, an issuer may be classified as Restricted or may be Under Review or Extended Review. In each case, investors should consider any investment opinion relating to such issuer (or its security and/or financial instruments) to be suspended or withdrawn and should not rely on the analyses and investment opinion(s) pertaining to such issuer (or its securities and/or financial instruments) nor should the analyses or opinion(s) be considered a solicitation of any kind. Sales persons and financial advisors affiliated with MLPF&S or any of its affiliates may not solicit purchases of securities or financial instruments that are Restricted or Under Review and may only solicit securities under Extended Review in accordance with firm policies.

Neither BofA Merrill Lynch nor any officer or employee of BofA Merrill Lynch accepts any liability whatsoever for any direct, indirect or consequential damages or losses arising from any use of this report or its contents.

Research Analysts

Economics

Jean-Michel Saliba
MENA Economist
MLI (UK)
+44 20 7995 8568
jean-michel.saliba@baml.com

MENA & FM equity

Hootan Yazhari, CFA >>
Research Analyst
Merrill Lynch (DIFC)
+971 4 4258218
hootan.yazhari@baml.com

Healthcare

Jamie Clark, CFA >>
Research Analyst
MLI (UK)
+44 20 7995 1300
jamie.clark@baml.com

Chemicals & Mining

Faisal AlAzmeh, CFA >>
Research Analyst
Merrill Lynch KSA Company
+966 11 299 3741
faisal.alazmeh@baml.com

Real Estate & Consumer

Abdelrali El Jattari >>
Research Analyst
Merrill Lynch (DIFC)
+971 4 4258231
abdelrali.eljattari@baml.com

Credit Research

Ali Dhaloomal
Research Analyst
MLI (UK)
+44 20 7996 9107
ali.dhaloomal@baml.com

Commodities

Francisco Blanch
Commodity & Deriv Strategist
MLPF&S
+1 646 855 6212
francisco.blanch@baml.com

Peter Helles
Commodity Strategist
MLI (UK)
+44 20 7996 8154
peter.helles@baml.com

Aerospace, Defence & Satellite Services

Celine Fornaro >>
Research Analyst
MLI (UK)
+44 20 7996 5515
celine.fornaro@baml.com

Benjamin Heelan >>
Research Analyst
MLI (UK)
+44 20 7996 5723
benjamin.heelan@baml.com

Energy

Anton Fedotov >>
Research Analyst
Merrill Lynch (Russia)
+7 495 662 6079
anton.a.fedotov@baml.com

Trading ideas and investment strategies discussed herein may give rise to significant risk and are not suitable for all investors. Investors should have experience in FX markets and the financial resources to absorb any losses arising from applying these ideas or strategies.
>> Employed by a non-US affiliate of MLPF&S and is not registered/qualified as a research analyst under the FINRA rules. Refer to "Other Important Disclosures" for information on certain BofA Merrill Lynch entities that take responsibility for this report in particular jurisdictions.

EXHIBIT 4

THE UNIVERSITY OF UTAH
S.J. QUINNEY COLLEGE OF LAW

PAUL G. CASSELL
Ronald N. Boyce Presidential Professor of Criminal Law
and University Distinguished Professor of Law
Telephone: 801-585-5202
cassellp@law.utah.edu

November 6, 2017

The Florida Bar
651 E. Jefferson St.
Tallahassee, FL 32399

Re: Bar Complaint By Mr. Dershowitz

Dear Florida Bar:

1. I have been advised of assertions made by Mr. Dershowitz to the Florida Bar concerning attorneys David Boies and Sigrid McCawley. I provide this affidavit at the request of counsel for Mr. Boies and Ms. McCawley in order to provide the Florida Bar with information that I have regarding the accuracy or inaccuracy of those assertions.

Civil Cases Against Epstein and Connections with Mr. Dershowitz

12. While the CVRA case was proceeding, Mr. Edwards and I also represented several of Epstein's victims in civil cases against Epstein. During those civil cases, Mr. Epstein was deposed about others who might have been involved in his sexual abuse of minor girls. Several civil attorneys asked Epstein about Mr. Dershowitz, and Epstein invoked his Fifth Amendment rights rather than answer questions about Dershowitz. Efforts were made to depose Dershowitz on his connections, but they were unsuccessful.

13. It was known at the time that Dershowitz had not only served as Epstein's legal counsel, but was also an extremely close personal friend of Epstein.[8] Household employees of Epstein also placed Dershowitz inside the Epstein mansion during time periods when Epstein was involved in sexually abusing minor girls. For example, household employee Juan Alessi testified under oath that one of the famous visitors who came to the Epstein mansion was "a very famous lawyer that I'm sure you know, Alan Dershowitz." (Alessi Depo., at 73:22-25.) According to Alessi's sworn testimony, Dershowitz came to the mansion "pretty often . . . at least four or five times a year" and would stay "two [or] three days." (73:22-25). Alessi also testified that Ms. Virginia Roberts Giuffre came to the house when Dershowitz was there. (73:18-20) Alessi answered "yes" when asked whether Dershowitz "had massages sometimes when he was there," (74:1-4). Alessi explained that "[a] massage was like a treat for everybody." (*Id.*)

14. Similar testimony was provided by another household employee, who worked for Epstein at a different time period -- Alfredo Rodriguez. According to Rodriguez, Dershowitz was at Epstein's mansion when underage girls were there to give massages. (Rodriguez Depo. at 78:13-25, 279:9-280:2). According to Rodriguez, Dershowitz stayed at the house in his role as Epstein's friend, as opposed to being his lawyer. (*Id.* at 279:5-8; 385:1-6).

15. Dershowitz was present alone, without his family, at the home of Jeffery Epstein in the presence of young girls. (*Id.* at 199:12-13, 279:9-12, 426:16-25, 427:1). Rodriguez testified that when the underage girls would come over, Dershowitz would drink wine and read books on the couch. (*Id.* at 426:16-25; 427:1.)

16. Ultimately Epstein choose to settle his civil suits with Jane Doe 1, Jane Doe 2, and Ms. Virginia Roberts Giuffre. However, Epstein did not settle the CVRA case, and Mr. Edwards and I have continued to attempt to secure factual evidence that would demonstrate that the plea decisions made by the Government were improper or improperly influenced.

---

[8] *See, e.g.*, Vicky Ward, *The Talented Mr. Epstein*, in *Vanity Fair* (Jan. 2005) (quoting Dershowitz as stating "I'm on my 20th book. . . . The only person outside of my immediate family that I send drafts to is Jeffrey."); *see also* Ray Gustini, *Vanity Fair Reminds Us When Jeffrey Epstein Wasn't a Creep*, in *The Wire* (June 21, 2011) (quoting Dershowitz as stating that even if Epstein went bankrupt, "I would be as interested in him as a friend if we had hamburgers on the boardwalk in Coney Island and talked about his ideas.").

(Excerpts)

Ms. Virginia Roberts Giuffre as a Witness in the Crime Victims Rights Act Case

17. One witness who became available to Mr. Edwards and me in the CVRA case was Ms. Virginia Roberts Giuffre. She had previously fled to Australia to escape from Mr. Epstein's sexual abuse. But at some point, in around spring 2014, we learned that she was in the United States and was willing to provide testimony in our CVRA case.

18. Because the CVRA case continues to be litigated today, nothing in this affidavit is intended to disclose – nor does it disclose – any confidential communications we received from Ms. Giuffre. However, I can report that in approximately May, 2014, I flew from Utah to Fort Lauderdale, Florida, where Mr. Edwards and I conducted a full-day interview of Ms. Giuffre to learn what evidence she had and to determine whether it was credible.

19. Without going into specific lawyer-client communications that we had with our client, Mr. Edwards and I determined that Ms. Giuffre's allegations of having been sexually abused by Epstein were credible. Ms. Giuffre also said that she had been sexually abused by Mr. Dershowitz. Mr. Edwards and I determined, based on that interview and subsequent careful follow-up, that her allegations were credible and were relevant to our allegations of improper influences affecting the Epstein plea negotiations.

20. Ms. Giuffre also credibly alleged that she had been sexually abused by other rich and powerful friends of Epstein. It appears that she had civil claims that she could file against Epstein's friends. However, because I am law professor with limited litigation resources, I recommended to Mr. Edwards that we should attempt to locate additional litigation support before pursuing lawsuits against abusers who could be expected to marshal considerable financial and other resources in response.

21. Mr. Edwards agreed and ultimately discussed the civil cases with Mr. David Boies, the name partner in the New York-based law firm of Boies Schiller Flexner LLP. I was aware that Mr. Boies not only had a well-regarded large litigation firm, but also had a strong track record of providing legal services for legal cases that might not otherwise attract attention from civil law firms (including cases involving sex trafficking victims).

22. I thought Mr. Boies was an excellent choice to help represent Ms. Giuffre. It is my understanding that Mr. Edwards conveyed to Mr. Boies – as well as, at some point, to his law partner Sigrid McCawley – that I had personally interviewed Ms. Giuffre, had investigated her allegations against Dershowitz and others, and had found them to be credible.

23. Because the CVRA revolved primarily around criminal law issues – an area where both Mr. Edwards and I had considerable experience – we did not require any assistance from Mr. Boies in the CVRA case and we did not approach him about assisting in that case. In 2014, we had successfully handled the CVRA case for more than six years, winning multiple motions without needing outside assistance, and anticipated no need for further help.

Filing Ms. Giuffre's Allegations and Dershowitz's Attacks

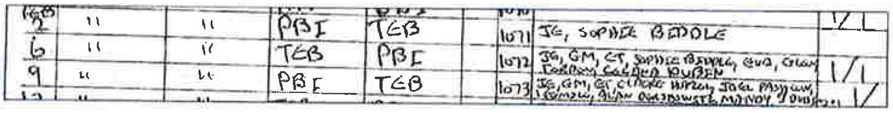
24. After more than six months of follow-up investigation, on December 30, 2014, Mr. Edwards and I filed Ms. Virginia Giuffre's allegations of abuse by Epstein, Dershowitz, and others in the CVRA case as part of a motion for her to join the CVRA case. The decision to file was made by Mr. Edwards and me. To the best of my recollection, I did not inform Mr. Boies of our intentions and, to my knowledge, neither he, nor his partner Ms. McCawley, nor attorney Stan Pottinger, were informed in advance of the date or contents of the filing.

33. In the course of those public attacks, Mr. Dershowitz made a number of statements to the press that were inconsistent with information that I possessed.

34. Dershowitz said, for example, that his wife accompanied him everywhere: "I've been married to the same woman for 28 years. She goes with me everywhere. People know that I won't argue a case or give a speech unless my wife travels with me." (*American Lawyer*, Jan. 15, 2015.) Dershowitz also denied being with Epstein in the company of young women: "I was never alone with Jeffrey Epstein in the company of any young women. I can tell you that unequivocally." (*Wall Street Journal*, Jan. 2, 2015). He cited flight manifests: "As far as the planes are concerned, there are flight manifests. They will prove I was never on any private airplane with any young women." (*Hala Gorani, CNN Live*, Jan. 5, 2015).

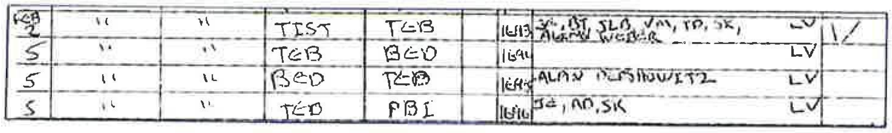
35. Contrary to these assertions, evidence I possessed showed the following things.

36. On February 9, 1998, Alan Dershowitz appears on flight #1073 to Teterboro Airport with Epstein, Ghislaine Maxwell, Emmy Taylor, Claire Hazel, Joel Pashcow, one unidentified female, Mandy and Dubins +1 (without his wife).



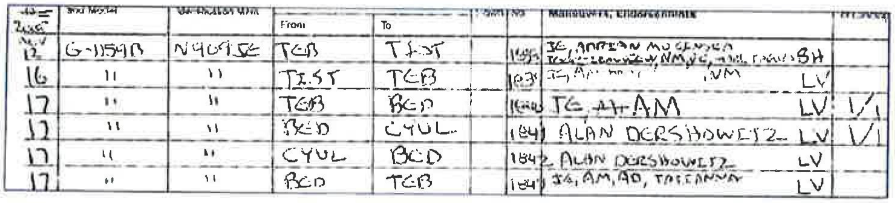
FEB 2	"	"	PBI	TEB		1071	JE, SOPHIE BIDDLE	1/1
6	"	"	TEB	PBI		1072	JE, GM, ET, SOPHIE BIDDLE, EVA, CLAIRE [illegible] RUBEN	1/1
9	"	"	PBI	TEB		1073	JE, GM, ET, CLAIRE HAZEL, JOEL PASHCOW, 1 FEMALE, ALAN DERSHOWITZ, MANDY, DUBINS+1	1/

37. On February 5, 2004, Alan Dershowitz appears on flight #1695 to Teterboro Airport, then boards flight #1696 to Palm Beach with Epstein and Sarah Kellen (without his wife).



FEB 2	"	"	TIST	TEB		1693	JE, BT, SLB, VM, TP, SK, ALAN WEBER	LV	1/
5	"	"	TEB	BED		1694		LV	
5	"	"	BED	TEB		1695	ALAN DERSHOWITZ	LV	
5	"	"	TEB	PBI		1696	JE, AD, SK	LV	

38. On November 17, 2005, Alan Dershowitz appears on flight #1841 to Montreal, Canada, then boards flight #1842 to Bedford, then #1843 to Teterboro with Epstein, Adriana Mucinska (AM), and Tatianna Kovylskna (without his wife).



2005	[illegible]	[illegible]	From	To	[illegible]	No.	Maneuvers, Endorsements		[illegible]
NOV 12	G-1159B	N909JE	TEB	TIST		1835	JE, ADRIANA MUCINSKA, [illegible], NM, JC, [illegible]	BH	
16	"	"	TIST	TEB		1838	JE, AM, [illegible], NM	LV	
17	"	"	TEB	BED		1840	JE, AM	LV	1/1
17	"	"	BED	CYUL		1841	ALAN DERSHOWITZ	LV	1/1
17	"	"	CYUL	BED		1842	ALAN DERSHOWITZ	LV	
17	"	"	BED	TEB		1843	JE, AM, AD, TATIANNA	LV	

39. To be clear, the flight logs that I had reviewed did not show Mr. Dershowitz alone with Ms. Giuffre. But I did not have a complete set of flight logs for the relevant time period and, in addition, it appears that Mr. Dershowitz had access to the flight logs before they were produced to law enforcement authorities. (*See* Police Detective Joe Recarey Depo. at 281, *Jane Doe No. 2 v. Epstein*, No. 9:08-cv-80119-KAM (Mar. 19, 2010) (Dershowitz personally provided the flight logs).)

40. To also be clear, the foregoing is not a complete recitation of the substantial information that I had that contradicted Mr. Dershowitz's claims.

43. During the lawsuit, Mr. Edwards and I realized that Ms. Giuffre would need separate legal counsel, apart from us, when she became a witness. As a result, we requested and helped to arrange for Ms. McCawley to provide that separate representation.

44. I recall that at some point during the lawsuit, Mr. Dershowitz met with Mr. Boies in an effort to persuade Mr. Boies of his innocence of abusing Ms. Giuffre. While I was not present at that meeting, it was my understanding from later events that Mr. Boies did not find Dershowitz's claims of innocence to be persuasive, particularly given that Dershowitz repeatedly made statements that were demonstrably untrue and apparently deliberate lies.

45. Ultimately, Mr. Edwards and I agreed to settle our defamation lawsuit against Mr. Dershowitz. The terms of the settlement were confidential and nothing in this affidavit should be construed as any disclosing confidential terms of that settlement.

46. That said, I can report that, contemporaneously with settling the case, my attorney, Jack Scarola, filed a Notice of Withdrawal of Motion for Partial Summary Judgment in which he responded to a public distortion made prominently by Mr. Dershowitz. Mr. Dershowitz claimed publicly that the word "mistake," contained in a jointly agreed public statement issued by him and us at the time of settlement, showed that we had supposedly exonerated him of the charges made by our client, Ms. Giuffre. To help keep the record from being distorted, Mr. Scarola included the following accurate statement:

> In the event that the noticed withdrawal is determined to be subject to Court approval, the Plaintiffs would show in support of this notice that Edwards and Cassell continue to represent Virginia Giuffre in separate pending matters, and shall continue to advance her legitimate legal interests in those matters. As expressly understood by the parties upon the execution of the Confidential Settlement Agreement and Mutual Release, Ms. Giuffre reaffirms her allegations, and the withdrawal of the referenced filings is not intended to be, and should not be construed as being, an acknowledgement by Edwards and Cassell that the allegations made by Ms. Giuffre were mistaken. Edwards and Cassell do acknowledge that the public filing in the Crime Victims' Rights Act case of their client's allegations against Defendant Dershowitz became a major distraction from the merits of the well-founded Crime Victims' Rights Act case by causing delay and, as a consequence, turned out to have been a tactical mistake. For that reason, Edwards and Cassell have chosen to withdraw the referenced filing as a condition of settlement.

Notice of Withdrawal of Motion for Partial Summary Judgment, *Edwards & Cassell v. Dershowitz*, No. 15-000072 (17th Judicial Circuit for Broward County, Florida) (Apr. 8, 2016).

No Extortion Plot Exists

53. I understand that Mr. Dershowitz has alleged that, in around fall 2014, Mr. Boies and Ms. McCawley were involved in an "extortion plot" to obtain money from Mr. Wexner. I am not aware of any such plot or even any suggestion of such a plot. I would also find any such suggestion entirely at odds with the professionalism that I have seen exhibited by Mr. Boies and Ms. McCawley in the course of professional dealings with them, including working closely with them as co-counsel to represent Ms. Giuffre in the complicated *Giuffre v. Maxwell* case. At all times they have exhibited nothing but the highest levels of professionalism and adherence to ethical requirements.

54. I expected no less since I knew that Mr. Boies had served as legal counsel in many high profile cases before I had the chance to work with him, including representing the Vice President of the United States before the United States Supreme Court in *Bush v. Gore*. And

while I did not know Ms. McCawley before our contact in this case, I have since learned of her tremendous devotion to her clients and her admirable motivation to insure that victims of sexual abuse receive effective legal representation when attacks on their credibility are made.

Conclusion

55. In light of all these facts, I believe that Mr. Dershowitz's bar complaint against Mr. Boies and Ms. McCawley is frivolous. The foregoing statements are my personal statements and are not an institutional position of the University of Utah. I request that this information be treated confidentially by the Florida Bar and not disseminated outside of the Bar (and those involved in this proceeding) in order to avoid any argument that I have somehow defamed anyone.

I declare that the foregoing statements are all true the best of my information and belief, and I make these statements under penalty of perjury.

Sincerely,

Paul G. Cassell, Esq.

Susan Baca
Notary Public
State of Utah
My Commission Expires April 25, 2021
696586

Susan Baca
Notary for the state of
Utah
Salt Lake County

From: Flipboard 10 for Today [editorialstaff@flipboard.com]
Sent: 7/13/2019 9:06:05 PM
To: jeevacation@gmail.com
Subject: 11 questions for Mueller

Flipboard

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Interesting stories worth your time.

The Seltzer Bubble

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

nytimes.com / SHEILA MARIKAR The drink of summer 2019 complies with just about every diet, befits all age groups, can be found in grocery stores across the United States (as has... read more

Here Are 11 Questions We’d Ask Robert Mueller

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

politico.com / POLITICO STAFF After two years, Congress finally has him on the stand. Here's our question list. In one week, Robert Mueller will raise his right hand on Capitol Hill,... read more

---

Mass layoffs: a history of cost cuts and psychological tolls

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

qz.com / Sarah Todd On July 8, Deutsche Bank began laying off 18,000 workers across its international offices as it retreats from the equities business, reducing the size... read more

---

These young scientists will shape the next 50 years of Moon research

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

nature.com / Alexandra Witze Half a century after Apollo, Nature profiles five researchers who are shaking up lunar exploration. Megha Bhatt grew up gazing at the Moon riding … read more

---

New White House Press Secretary 'Likes A Challenge'

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

npr.org / Tamara Keith Years before she would become President Trump's third press secretary, Stephanie Grisham had a photo of the White House hanging on her office wall in... read more

---

What the Era of 'Big Plans' Can Teach Us Today

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

citylab.com / Ezra Haber Glenn A new exhibition looks back at the period of grand urban design and social reform in New York City, Boston, and Chicago. This summer, visitors to... read more

---

True Story of JFK Jr. and Carolyn Bessette's Last Days Before the 1999 Plane Crash

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

townandcountrymag.com / Adrienne Gaffney In early May 1999, George magazine hosted a table at the White House Correspondents Dinner. The glossy’s guest list reflected the template for the… read more

For thousands of fans, queuing is a Wimbledon tradition

apnews.com / MATTIAS KAREN WIMBLEDON, England (AP) — For Viv Kean, and thousands of tennis fans like her, the Wimbledon experience always starts in a tent. In a small park across… read more

---

OK, No Burgers—But What Should You Grill?

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

+

bloomberg.com / Kate Krader Last week, Bloomberg Pursuits “accidentally” disrupted Independence Day by announcing that burgers should not be grilled. That advice came from three… read more

---

SI’s 24 USWNT Commemorative Digital Covers

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

si.com / Avi Creditor Four years ago when the U.S. women won the World Cup, they did it on a magazine dark week for Sports Illustrated. With no issue to print, we ... read more

---

Your guide to movies, music, books and more.

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

SEE YOUR FLIPBOARD ON THE WEB
OR DOWNLOAD THE MOBILE APP

You have received this email because you signed up for an account on Flipboard with the email address: jeevacation@gmail.com.

UNSUBSCRIBE | PRIVACY POLICY

©2019 Flipboard, Inc. Made with love at 735 Emerson Street, Palo Alto, CA 94301

From: Richard Kahn [redacted]
Sent: 5/9/2018 1:51:01 AM
To: Jeffrey Epstein [jeevacation@gmail.com]
Subject: AT&T confirms it paid Trump lawyer Michael Cohen for 'insights' on administration

Importance: High

https://www.cnbc.com/2018/05/08/att-confirms-it-paid-trump-lawyer-michael-cohen-for-insights-on-administration.html

AT&T confirms it paid Trump lawyer Michael Cohen for 'insights' on administration

11:57 AM ET Fri, 4 May 2018 | 01:18

Getty Images

Michael Cohen, longtime personal lawyer and confidante for President Donald Trump, arrives at the United States District Court Southern District of New York on April 26, 2018 in New York City.

Telecommunications giant AT&T said Tuesday night that it had paid President Donald Trump's lawyer Michael Cohen for "insights" about the Trump administration.

AT&T's admission came after a lawyer for porn star Stormy Daniels claimed the company, drug giant Novartis and a Russian oligarch had all made payments to Cohen's shell company.

Daniels' lawyer, Michael Avenatti, said AT&T had made four separate payments of $50,000 apiece to Cohen's company, for a total of $200,000 in late 2017 and into early 2018.

That company, Essential Consultants, was created by Cohen in October 2016 and soon after was used to make a $130,000 hush-money payment to Daniels.

In a prepared statement to CNBC, AT&T said Cohen's company "was one of several firms we engaged in early 2017 to provide insights into understanding the new administration."

"They did no legal or lobbying work for us, and the contract ended in December 2017," AT&T said.

The company did not say how much it had paid Cohen, who was the president's personal lawyer at the time.

AT&T is in the midst of pursuing an $85 billion acquisition of Time Warner. The U.S. Justice Department has sued to block that deal.

In a report on Cohen's company, Avenatti's law firm said that Novartis in late 2017 and early 2018 made four separate payments to Essential Consultants totaling nearly $400,000.

"Following these payments, reports surfaced that Mr. Trump took a dinner with the incoming CEO of Novartis before Mr. Trump's speech at the World Economic Forum in Davos, Switzerland in late January 2018," Avenatti's report said.

That CEO, Vas Narasimhan, was joined with a group of other companies' executives at that dinner.

A Novartis spokesperson said in a statement that "any agreements with Essential Consultants were entered before our current CEO taking office in February of this year and have expired."

Avenatti's report says another company, Korea Aerospace Industries LTD, paid Essential Consultants $150,000 in November 2017.

Avenatti's client Daniels, whose real name is Stephanie Clifford, was paid $130,000 by Essential Consultants on the eve of the 2016 presidential election.

Daniels says the money was in exchange for her signing a deal that required her to remain silent about an affair she claims to have had with Trump in 2006, shortly after the birth of his youngest son.

The White House has denied that Trump had sex with the adult film actress.

Neither the White House nor Cohen had an immediate comment on Avenatti's new allegations about payments to Cohen's company.

WATCH: Feds saw log of Cohen's phone calls

Richard Kahn
HBRK Associates Inc.
575 Lexington Avenue, 4th Floor
New York, NY 10022
[redacted]

Penguin Press

National Pub date: February 19, 2019
Title: *DEEP THINKING*
Subtitle: *Twenty-Five Ways of Looking at AI*
By: John Brockman
Length: 90,000 words

Headline: Science world luminary John Brockman assembles twenty-five of the most important scientific minds, people who have been thinking about the field artificial intelligence for most of their careers for an unparalleled round-table examination about mind, thinking, intelligence and what it means to be human.

Description:

"Artificial intelligence is today's story—the story behind all other stories. It is the Second Coming and the Apocalypse at the same time: Good AI versus evil AI." —John Brockman

More than sixty years ago, mathematician-philosopher Norbert Wiener published a book on the place of machines in society that ended with a warning: "we shall never receive the right answers to our questions unless we ask the right questions…. The hour is very late, and the choice of good and evil knocks at our door."

In the wake of advances in unsupervised, self-improving machine learning, a small but influential community of thinkers is considering Wiener's words again. In *Deep Thinking*, John Brockman gathers their disparate visions of where AI might be taking us.

The fruit of the long history of Brockman's profound engagement with the most important scientific minds who have been thinking about AI—from Alison Gopnik and David Deutsch to Frank Wilczek and Stephen Wolfram— *Deep Thinking* is an ideal introduction to the landscape of crucial issues AI presents.

The collision between opposing perspectives is salutary and exhilarating; some of these figures, such as computer scientist Stuart Russell, Skype co-founder Jaan Tallinn, and physicist Max Tegmark, are deeply concerned with the threat of AI, including the existential one, while others, notably robotics entrepreneur Rodney Brooks, philosopher Daniel Dennett, and bestselling author Steven Pinker, have a very different view. Serious, searching and authoritative, *Deep Thinking* lays out the intellectual landscape of one of the most important topics of our time.

Participants in The Deep Thinking Project

Chris Anderson is an entrepreneur; a roboticist; former editor-in-chief of *Wired*; co-founder and CEO of 3DR; and author of *The Long Tail, Free,* and *Makers.*

Rodney Brooks is a computer scientist; Panasonic Professor of Robotics, emeritus, MIT; former director, MIT Computer Science Lab; and founder, chairman, and CTO of Rethink Robotics. He is the author of *Flesh and Machines.*

George M. Church is Robert Winthrop Professor of Genetics at Harvard Medical School; Professor of Health Sciences and Technology, Harvard-MIT; and co-author (with Ed Regis) of *Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves.*

Daniel C. Dennett is University Professor and Austin B. Fletcher Professor of Philosophy and director of the Center for Cognitive Studies at Tufts University. He is the author of a dozen books, including *Consciousness Explained* and, most recently, From *Bacteria to Bach and Back: The Evolution of Minds.*

David Deutsch is a quantum physicist and a member of the Centre for Quantum Computation at the Clarendon Laboratory, Oxford University. He is the author of *The Fabric of Reality* and *The Beginning of Infinity.*

Anca Dragan is an assistant professor in the Department of Electrical Engineering and Computer Sciences at UC Berkeley. She co-founded and serves on the steering committee for the Berkeley AI Research (BAIR) Lab and is a co-principal investigator in Berkeley's Center for Human-Compatible AI.

George Dyson is a historian of science and technology and the author of *Baidarka: the Kayak, Darwin Among the Machines, Project Orion,* and *Turing's Cathedral.*

Peter Galison is a science historian, Joseph Pellegrino University Professor and co-founder of
the Black Hole Initiative at Harvard University, and the author of *Einstein's Clocks and Poincaré's Maps: Empires of Time.*

Neil Gershenfeld is a physicist and director of MIT's Center for Bits and Atoms. He is the author of *FAB*, co-author (with Alan Gershenfeld & Joel Cutcher-Gershenfeld) of *Designing Reality*, and founder of the global fab lab network.

Alison Gopnik is a developmental psychologist at UC Berkeley; her books include The Philosophical Baby and, most recently, *The Gardener and the Carpenter: What the New Science of Child Development Tells Us About the Relationship Between Parents and Children.*

Tom Griffiths is Henry R. Luce Professor of Information, Technology, Consciousness, and Culture at Princeton University. He is co-author (with Brian Christian) of *Algorithms to Live By.*

W. Daniel "Danny" Hillis is an inventor, entrepreneur, and computer scientist, Judge Widney Professor of Engineering and Medicine at USC, and author of *The Pattern on the Stone: The Simple Ideas That Make Computers Work.*

Caroline A. Jones is a professor of art history in the Department of Architecture at MIT and author of *Eyesight Alone: Clement Greenberg's Modernism and the Bureaucratization of the Senses; Machine in the Studio: Constructing the Postwar American Artist;* and *The Global Work of Art.*

David Kaiser is Germeshausen Professor of the History of Science and professor of physics at MIT, and head of its Program in Science, Technology & Society. He is the author of *How the Hippies Saved Physics: Science, Counterculture*, and the *Quantum Revival and American Physics and the Cold War Bubble* (forthcoming).

Seth Lloyd is a theoretical physicist at MIT, Nam P. Suh Professor in the Department of Mechanical Engineering, and an external professor at the Santa Fe Institute. He is the author of *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos.*

Hans Ulrich Obrist is artistic director of the Serpentine Gallery, London, and the author of *Ways of Curating* and *Lives of the Artists, Lives of the Architects.*

Judea Pearl is professor of computer science and director of the Cognitive Systems Laboratory at UCLA. His most recent book, co-authored with Dana Mackenzie, is *The Book of Why: The*

Alex "Sandy" Pentland is Toshiba Professor and professor of media arts and sciences, MIT; director of the Human Dynamics and Connection Science labs and the Media Lab Entrepreneurship Program, and the author of *Social Physics.*
New Science of Cause and Effect.

Steven Pinker, a Johnstone Family Professor in the Department of Psychology at Harvard University, is an experimental psychologist who conducts research in visual cognition, psycholinguistics, and social relations. He is the author of eleven books, including *The Blank Slate, The Better Angels of Our Nature*, and, most recently, *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress.*

Venki Ramakrishnan is a scientist at the Medical Research Council Laboratory of Molecular Biology, Cambridge University; recipient of the Nobel Prize in Chemistry (2009); current president of the Royal Society; and the author of *Gene Machine: The Race to Discover the Secrets of the Ribosome.*

Stuart Russell is a professor of computer science and Smith-Zadeh Professor in Engineering at UC Berkeley. He is the coauthor (with Peter Norvig) of *Artificial Intelligence: A Modern Approach.*

Jaan Tallin, a computer programmer, theoretical physicist, and investor, is a co-developer of Skype and Kazaa.

Max Tegmark is an MIT physicist and AI researcher; president of the Future of Life Institute; scientific director of the Foundational Questions Institute; and the author of *Our Mathematical Universe* and *Life 3.0: Being Human in the Age of Artificial Intelligence.*

Frank Wilczek is Herman Feshbach Professor of Physics at MIT, recipient of the 2004 Nobel Prize in physics, and the author of *A Beautiful Question: Finding Nature's Deep Design*.

Stephen Wolfram is a scientist, inventor, and the founder and CEO of Wolfram Research. He is the creator of the symbolic computation program Mathematica and its programming language, Wolfram Language, as well as the knowledge engine Wolfram|Alpha. He is also the author of *A New Kind of Science.*

Deep Thinking
Twenty-five Ways of Looking at AI

edited by John Brockma

Penguin Press — February 19, 2019

Table of Contents

Acknowledgments

Introduction: On the Promise and Peril of AI
by John Brockman
Artificial intelligence is today's story—the story behind all other stories. It is the Second Coming and the Apocalypse at the same time: Good AI versus evil AI. This book comes out of an ongoing conversation with a number of important thinkers, both in the world of AI and beyond it, about what AI is and what it means. Called the Deep Thinking Project, this conversation began in earnest in September 2016, in a meeting at the Mayflower Grace Hotel in Washington, Connecticut with some of the book's contributors.

What quickly emerged from that first meeting is that the excitement and fear in the wider culture surrounding AI now has an analogue in the way Norbert Wiener's ideas regarding "cybernetics" worked their way through the culture, particularly in the 1960's, as artists began to incorporate thinking about new technologies into their work. I witnessed the impact of those ideas at close hand; indeed it's not too much to say they set me off on my life's path. With the advent of the digital era beginning in the early 1970s, people stopped talking about Wiener, but today, his Cybernetic Idea has been so widely adopted that it's internalized to the point where it no longer needs a name. It's everywhere, it's in the air, and it's a fitting a place to begin.

Seth Lloyd: Wrong, but More Relevant Than Ever
It is exactly in the extension of the cybernetic idea to human beings that Wiener's conceptions missed their target.

Judea Pearl: The Limitations of Opaque Learning Machines
Deep learning has its own dynamics, it does its own repair and its own optimization, and it gives you the right results most of the time. But when it doesn't, you don't have a clue about what went wrong and what should be fixed.

Stuart Russell: The Purpose Put Into the Machine
We may face the prospect of superintelligent machines—their actions by definition unpredictable by us and their imperfectly specified objectives conflicting with our own—whose motivation to preserve their existence in order to achieve those objectives may be insuperable.

George Dyson: The Third Law
Any system simple enough to be understandable will not be complicated enough to behave intelligently, while any system complicated enough to behave intelligently will be too complicated to understand.

Daniel C. Dennett: What Can We Do?
We don't need artificial conscious agents. We need intelligent tools.

Rodney Brooks: The Inhuman Mess Our Machines Have Gotten Us Into
We are in a much more complex situation today than Wiener foresaw, and I am worried that it is much more pernicious than even his worst imagined fears.

Frank Wilczek: The Unity of Intelligence
The advantages of artificial over natural intelligence appear permanent, while the advantages of natural over artificial intelligence, though substantial at present, appear transient.

Max Tegmark: Let's Aspire to More Than Making Ourselves Obsolete
We should analyze what could go wrong with AI to ensure that it goes right.

Jaan Tallinn: Dissident Messages
Continued progress in AI can precipitate a change of cosmic proportions—a runaway process that will likely kill everyone.

Steven Pinker: Tech Prophecy and the Underappreciated Causal Power of Ideas
There is no law of complex systems that says that intelligent agents must turn into ruthless megalomaniacs.

David Deutsch: Beyond Reward and Punishment
Misconceptions about human thinking and human origins are causing corresponding misconceptions about AGI and how it might be created.

Tom Griffiths: The Artificial Use of Human Beings
Automated intelligent systems that will make good inferences about what people want must have good generative models for human behavior.

Anca Dragan: Putting the Human into the AI Equation
In the real world, an AI must interact with people and reason about them. People will have to formally enter the AI problem definition somewhere.

Chris Anderson: Gradient Descent
Just because AI systems sometimes end up in local minima, don't conclude that this makes them any less like life. Humans—indeed, probably all life-forms—are often stuck in local minima.

David Kaiser: "Information" for Wiener, for Shannon, and for Us
Many of the central arguments in The Human Use of Human Beings *seem closer to the 19th century than the 21st. Wiener seems not to have fully embraced Shannon's notion of information as consisting of irreducible, meaning-free bits.*

Neil Gershenfeld: Scaling
Although machine making and machine thinking might appear to be unrelated trends, they lie in each other's futures.

W. Daniel Hillis: The First Machine Intelligences
Hybrid superintelligences such as nation states and corporations have their own emergent goals and their actions are not always aligned to the interests of the people who created them.

Venki Ramakrishnan: Will Computers Become Our Overlords?
Our fears about AI reflect the belief that our intelligence is what makes us special.

Alex "Sandy" Pentland: The Human Strategy
How can we make a good human-artificial ecosystem, something that's not a machine society but a cyberculture in which we can all live as humans—a culture with a human feel to it?

Hans Ulrich Obrist: Making the Invisible Visible: Art Meets AI
Many contemporary artists are articulating various doubts about the promises of AI and reminding us not to associate the term "artificial intelligence" solely with positive outcomes.

Alison Gopnik: AIs versus Four-Year-Olds
Looking at what children do may give programmers useful hints about directions for computer learning.

Peter Galison: Algorists Dream of Objectivity
By now, the legal, ethical, formal, and economic dimensions of algorithms are all quasi-infinite.

George M. Church: The Rights of Machines
Probably we should be less concerned about us-versus-them and more concerned about the rights of all sentients in the face of an emerging unprecedented diversity of minds.

Caroline A. Jones: The Artistic Use of Cybernetic Beings
The work of cybernetically inclined artists concerns the emergent behaviors of life that elude AI in its current condition.

Stephen Wolfram: Artificial Intelligence and the Future of Civilization
The most dramatic discontinuity will surely be when we achieve effective human immortality. Whether this will be achieved biologically or digitally isn't clear, but inevitably it will be achieved.

Introduction: On the Promise and Peril of AI

John Brockman

Artificial intelligence is today's story—the story behind all other stories. It is the Second Coming and the Apocalypse at the same time: Good AI versus evil AI. This book comes out of an ongoing conversation with a number of important thinkers, both in the world of AI and beyond it, about what AI is and what it means. Called the Deep Thinking Project, this conversation began in earnest in September 2016, in a meeting at the Mayflower Grace Hotel in Washington, Connecticut with some of the book's contributors.

What quickly emerged from that first meeting is that the excitement and fear in the wider culture surrounding AI now has an analogue in the way Norbert Wiener's ideas regarding "cybernetics" worked their way through the culture, particularly in the 1960's, as artists began to incorporate thinking about new technologies into their work. I witnessed the impact of those ideas at close hand; indeed it's not too much to say they set me off on my life's path. With the advent of the digital era beginning in the early 1970s, people stopped talking about Wiener, but today, his Cybernetic Idea has been so widely adopted that it's internalized to the point where it no longer needs a name. It's everywhere, it's in the air, and it's a fitting a place to begin.

New Technologies=New Perceptions

Before AI, there was Cybernetics—the idea of automatic, self-regulating control, laid out in Norbert Wiener's foundational text of 1948. I can date my own serious exposure to it to 1966, when the composer John Cage invited me and four or five other young arts people to join him for a series of dinners—an ongoing seminar about media, communications, art, music, and philosophy that focused on Cage's interest in the ideas of Wiener, Claude Shannon, and Marshall McLuhan, all of whom had currency in the New York art circles in which I was then moving. In particular, Cage had picked up on McLuhan's idea that by inventing electronic technologies we had externalized our central nervous system—that is, our minds—and that we now had to presume that "there's only one mind, the one we all share."

Ideas of this nature were beginning to be of great interest to the artists I was working with in New York at the Film-Makers' Cinémathèque, where I was program manager for a series of multimedia productions called the New Cinema 1 (also known as the Expanded Cinema Festival), under the auspices of avant-garde filmmaker and impresario Jonas Mekas. They included visual artists Claes Oldenburg, Robert Rauschenberg, Andy Warhol, Robert Whitman; kinetic artists Charlotte Moorman and Nam June Paik; happenings artists Allan Kaprow and Carolee Schneemann; dancer Tricia Brown; filmmakers Jack Smith, Stan Vanderbeek, Ed Emshwiller, and the Kuchar brothers; avant-garde dramatist Ken Dewey; poet Gerd Stern and the USCO group; minimalist musicians Lamonte Young and Terry Riley; and through Warhol, the music group, The Velvet Underground. Many of these people were reading Wiener, and cybernetics was in the air. It was at one of these dinners that Cage reached into his briefcase and took out a copy of *Cybernetics* and handed it to me, saying, "This is for you."

During the Festival, I received an unexpected phone call from Wiener's colleague Arthur K. Solomon, head of Harvard's graduate program in biophysics. Wiener had died the year before, and Solomon and Wiener's other close colleagues at MIT and Harvard had been reading about the Expanded Cinema Festival in the *New York Times* and were intrigued by the connection to Wiener's work. Solomon invited me to bring some of the artists up to Cambridge to meet with him and a group that included MIT sensory-communications researcher Walter Rosenblith, Harvard applied mathematician Anthony Oettinger, and MIT engineer Harold "Doc" Edgerton, inventor of the strobe light.

Like many other "art meets science" situations I've been involved in since, the two-day event was an informed failure: ships passing in the night. But I took it all onboard and the event was consequential in some interesting ways—one of which came from the fact that they took us to see "the" computer. Computers were a rarity back then; at least, none of us on the visit had ever seen one. We were ushered into a large space on the MIT campus, in the middle of which there was a "cold room" raised off the floor and enclosed in glass, in which technicians wearing white lab coats, scarves, and gloves were busy collating punch cards coming through an enormous machine. When I approached, the steam from my breath fogged up the window into the cold room. Wiping it off, I saw "the" computer. I fell in love.

Later, in the Fall of 1967, I went to Menlo Park to spend time with Stewart Brand, whom I had met in New York in 1965 when he was a satellite member of the USCO group of artists. Now, with his wife Lois, a mathematician, he was preparing the first edition of *The Whole Earth Catalog* for publication. While Lois and the team did the heavy lifting on the final mechanicals for WEC, Stewart and I sat together in a corner for two days, reading, underlining, and annotating the same paperback copy of *Cybernetics* that Cage had handed to me the year before, and debating Wiener's ideas.

Inspired by this set of ideas, I began to develop a theme, a mantra of sorts, that has informed my endeavors since: "new technologies = new perceptions." Inspired by communications theorist Marshall McLuhan, architect-designer Buckminster Fuller, futurist John McHale, and cultural anthropologists Edward T. (Ned) Hall and Edmund Carpenter, I started reading avidly in the field of information theory, cybernetics, and systems theory. McLuhan suggested I read biologist J.Z. Young's *Doubt and Certainty in Science* in which he said that we create tools and we mold ourselves through our use of them. The other text he recommended was Warren Weaver and Claude Shannon's 1949 paper "Recent Contributions to the Mathematical Theory of Communication," which begins: "The word *communication* will be used here in a very broad sense to include all of the procedures by which one mind may affect another. This, of course, involves not only written and oral speech, but also music, the pictorial arts, the theater, the ballet, and in fact all human behavior."

Who knew that within two decades of that moment we would begin to recognize the brain as a computer? And in the next two decades, as we built our computers into the Internet, that we would begin to realize that the brain is not a computer, but a network of computers? Certainly not Wiener, a specialist in analogue feedback circuits designed to control machines, nor the artists, nor, least of all, myself.

"We must cease to kiss the whip that lashes us."

Two years after *Cybernetics,* in 1950, Norbert Wiener published *The Human Use of Human Beings*—a deeper story, in which he expressed his concerns about the runaway commercial exploitation and other unforeseen consequences of the new technologies of control. I didn't read *The Human Use of Human Beings* until the spring of 2016, when I picked up my copy, a first edition, which was sitting in my library next to *Cybernetics.* What shocked me was the realization of just how prescient Wiener was in 1950 about what's going on today. Although the first edition was a major bestseller—and, indeed, jump-started an important conversation—under pressure from his peers Wiener brought out a revised and milder edition in 1954, from which the original concluding chapter, "Voices of Rigidity," is conspicuously absent.

Science historian George Dyson points out that in this long-forgotten first edition, Wiener predicted the possibility of a "threatening new Fascism dependent on the *machine à gouverner*":

> No elite escaped his criticism, from the Marxists and the Jesuits ("all of Catholicism is indeed essentially a totalitarian religion") to the FBI ("our great merchant princes have looked upon the propaganda technique of the Russians, and have found that it is good") and the financiers lending their support "to make American capitalism and the fifth freedom of the businessman supreme throughout the world." Scientists . . . received the same scrutiny given the Church: "Indeed, the heads of great laboratories are very much like Bishops, with their association with the powerful in all walks of life, and the dangers they incur of the carnal sins of pride and of lust for power."

This jeremiad did not go well for Wiener. As Dyson puts it:

> These alarms were discounted at the time, not because Wiener was wrong about digital computing but because larger threats were looming as he completed his manuscript in the fall of 1949. Wiener had nothing against digital computing but was strongly opposed to nuclear weapons and refused to join those who were building digital computers to move forward on the thousand-times-more-powerful hydrogen bomb.

Since the original of *The Human Use of Human Beings* is now out of print, lost to us is Wiener's *cri de coeur*, more relevant today than when he wrote it, sixty-eight years ago: "We must cease to kiss the whip that lashes us."

Mind, Thinking, Intelligence

Among the reasons we don't hear much about "Cybernetics" today, two are central: First, although *The Human Use of Human Beings* was considered an important book in its time, it ran counter to the aspirations of many of Wiener's colleagues, including John von Neumann and Claude Shannon, who were interested in the commercialization of the new technologies. Second, computer pioneer John McCarthy disliked Wiener and refused to use Wiener's term "Cybernetics." McCarthy, in turn, coined the term "artificial intelligence" and became a founding father of that field.

As Judea Pearl, who, in the 1980s, introduced a new approach to artificial intelligence called Bayesian networks, explained to me:

> What Wiener created was excitement to believe that one day we are going to make an intelligent machine. He wasn't a computer scientist. He talked feedback, he talked communication, he talked analog. His working metaphor was a feedback circuit, which he was an expert in. By the time the digital age began in the early 1960s people wanted to talk programming, talk codes, talk about computational functions, talk about short-term memory, long-term memory—meaningful computer metaphors. Wiener wasn't part of that, and he didn't reach the new generation that germinated with his ideas. His metaphors were too old, *passé*. There were new means already available that were ready to capture the human imagination." By 1970, people were no longer talking about Wiener.

One critical factor missing in Wiener's vision was the cognitive element: mind, thinking, intelligence. As early as 1942, at the first of a series of foundational interdisciplinary meetings about the control of complex systems that would come to be known as the Macy conferences, leading researchers were arguing for the inclusion of the cognitive element into the conversation. While von Neumann, Shannon, and Wiener were concerned about systems of control and communication of observed systems, Warren McCullough wanted to include mind. He turned to cultural anthropologists Gregory Bateson and Margaret Mead to make the connection to the social sciences. Bateson in particular was increasingly talking about patterns and processes, or "the pattern that connects." He called for a new kind of systems ecology in which organisms and the environment in which they live are one in the same, and should be considered as a single circuit. By the early 1970s the Cybernetics of observed systems—1st order Cybernetics—moved to the Cybernetics of observing systems—2nd order Cybernetics—or "the Cybernetics of Cybernetics", as coined by Heinz von Foerster, who joined the Macy conferences in the mid 1950s, and spearheaded the new movement.

Cybernetics, rather than disappearing, was becoming metabolized into *everything*, so we no longer saw it as a separate, distinct new discipline. And there it remains, hiding in plain sight.

"The Shtick of the Steins"

My own writing about these issues at the time was on the radar screen of the 2nd order Cybernetics crowd, including Heinz von Foerster as well as John Lilly and Alan Watts, who were the co-organizers of something called "The AUM Conference," shorthand for "The American University of Masters", which took place in Big Sur in 1973, a gathering of philosophers, psychologists, and scientists, each of whom asked to lecture on his own work in terms of its relationship to the ideas of British mathematician G. Spencer Brown presented in his book, *Laws of Form*.

I was a bit puzzled when I received an invitation—a very late invitation indeed—which they explained was based on their interest in the ideas I presented in a book called *Afterwords*, which were very much on their wavelength. I jumped at the opportunity, the main reason being that the keynote speaker was none other than Richard Feynman. I love to spend time with physicists, the reason being that they think about the universe, i.e. everything. And no physicist was reputed to be articulate as Feynman. I couldn't wait to meet him. I accepted. That said, I am not a scientist, and I had never entertained the idea of getting on a stage and delivering a "lecture" of any kind, least of all a commentary on an obscure mathematical theory in front of a group identified as the world's most interesting thinkers. Only upon my arrival in Big Sur did I find out the reason for my very late invitation. "When is Feynman's talk?" I asked at the desk. "Oh, didn't Alan Watts tell you? Richard is ill and has been hospitalized. You're his replacement. And, by the way, what's the title of your keynote lecture?"

I tried to make myself invisible for several days. Alan Watts, realizing that I was avoiding the podium, woke me up one night with a 3am knock on the door of my room. I opened the door to find him standing in front of me wearing a monk's robe with a hood that covering much of his face. His arms extended, he held a lantern in one hand, and a magnum of scotch on the other.

"John", he said in a deep voice with a rich aristocratic British accent, "you are a phony." "And, John", he continued, I am a phony. But John, I am a *real* phony!"

The next day I gave my lecture, entitled "Einstein, Gertrude Stein, Wittgenstein, and Frankenstein." Einstein: the revolution in 20th century physics; Gertrude Stein: the first writer who made integral to her work the idea of an indeterminate and discontinuous universe. Words represented neither character nor activity: A rose is a rose is a rose, and a universe is a universe is a universe.); Wittgenstein: the world as limits of language. "The limits of my language mean the limits of my world". The end of the distinction between observer and observed. Frankenstein: Cybernetics AI, robotics, all the essayists in this volume.

The lecture had unanticipated consequences. Among the participants at the AUM Conference were several authors of #1 New York Times bestsellers, yet no one there had a literary agent. And I realized that all were engaged in writing a genre of book both unnamed and unrecognized by New York publishers. Since I had an MBA from Colombia Business School, and a series of relative successes in business, I was dragooned into becoming an agent, initially for Gregory Bateson and John Lilly, whose books I sold quickly, and for sums that caught my attention, thus kick-starting my career as a literary agent.

I never did meet Richard Feynman.

The Long AI Winters

This new career put me in close touch with most of the AI pioneers, and over the decades I rode with them on waves of enthusiasm, and into valleys of disappointment.

In the early '80s the Japanese government mounted a national effort to advance AI. They called it the 5th Generation; their goal was to change the architecture of computation by breaking "the von Neumann bottleneck", by creating a massively parallel computer. In so doing, they hoped to jumpstart their economy and become a dominant world power in the field. In1983, the leader of the Japanese 5th Generation consortium came to New York for a meeting organized by Heinz Pagels, the president of the New York Academy of Sciences. I had a seat at the table alongside the leaders of the 1st generation, Marvin Minsky and John McCarthy, the 2nd generation, Edward Feigenbaum and Roger Schank, and Joseph Traub, head of the National Supercomputer Consortium.

In 1981 with Heinz's help, I had founded "The Reality Club" (the precursor to the non-profit Edge.org), whose initial interdisciplinary meetings took place in the Board Room at the NYAS. Heinz was working on his book, *Dreams of Reason: The Rise of the Science of Complexity*, which he considered to be a research agenda for science in the 1990's.

Through the Reality Club meetings, I got to know two young researchers who were about to play key roles in revolutionizing computer science. At MIT in the late seventies, Danny Hillis developed the algorithms that made possible the massively parallel computer. In 1983, his company, Thinking Machines, built the world's fastest supercomputer by utilizing parallel architecture. His "connection machine," closely reflected the workings of the human mind. Seth Lloyd at Rockefeller University was undertaking seminal work in the fields of quantum computation and quantum communications, including proposing the first technologically feasible design for a quantum computer.

And the Japanese? Their foray into artificial intelligence failed, and was followed by twenty years of anemic economic growth. But, the leading US scientists took this program very seriously. And Feigenbaum, who was the cutting-edge computer scientist of the day, teamed up with McCorduck to write a book on these developments. *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World* was published in 1983. We had a code name for the project: "It's coming, it's coming!" But it didn't come; it went.

From that point on I've worked with researchers in nearly every variety of AI and complexity, including Rodney Brooks, Hans Moravec, John Archibald Wheeler, Benoit Mandelbrot, John Henry Holland, Danny Hillis, Freeman Dyson, Chris Langton, Doyne Farmer, Geoffrey West, Stuart Russell, and Judea Pearl.

An Ongoing Dynamical Emergent System

From the initial meeting in Washington, CT to the present, I arranged a number of dinners and discussions in London and Cambridge, Massachusetts, as well as a public event at London's City Hall. Among the attendees were distinguished scientists, science historians, and communications theorists, all of whom have been thinking seriously about AI issues for their entire careers.

I commissioned essays from a wide range of contributors, with or without references to Wiener (leaving it up to each participant). In the end, 25 people wrote essays, all individuals concerned about what is happening today in the age of AI. *Deep Thinking* in not *my* book, rather it is *our* book: Seth Lloyd, Judea Pearl, Stuart Russell, George Dyson, Daniel C. Dennett, Rodney Brooks, Frank Wilczek, Max Tegmark, Jaan Tallinn, Steven Pinker, David Deutsch, Tom Griffiths, Anca Dragan, Chris Anderson, David Kaiser, Neil Gershenfeld, W. Daniel Hillis, Venki Ramakrishnan, Alex "Sandy" Pentland, Hans Ulrich Obrist, Alison Gopnik, Peter Galison, George M. Church, Caroline A. Jones, Stephen Wolfram.

I see The Deep Thinking Project as an ongoing dynamical emergent system, a presentation of the ideas of a community of sophisticated thinkers who are bringing their experience and erudition to bear in challenging the prevailing digital AI narrative as they communicate their thoughts to one another. The aim is to present a mosaic of views which will help make sense out of this rapidly emerging field.

I asked the essayists to consider:

(a) The Zen-like poem "Thirteen Ways of Looking at a Blackbird," by Wallace Stevens, which he insisted was "not meant to be a collection of epigrams or of ideas, but of sensations." It is an exercise in "perspectivism," consisting of short, separate sections, each of which mentions blackbirds in some way. The poem is about his own imagination; it concerns what he attends to.

(b) The parable of the blind men and an elephant. Like the elephant, AI is too big a topic for any one perspective, never mind the fact that no two people seem to see things the same way.

What do we want the book to do? Stewart Brand has noted that "revisiting pioneer thinking is perpetually useful. And it gives a long perspective that invites thinking in decades and centuries about the subject. All contemporary discussion, is bound to age badly and immediately without the longer perspective."

Danny Hillis wants people in AI to realize how they've been programmed by Wiener's book. "You're executing its road map," he says, and you just don't realize it."

Dan Dennett would like to "let Wiener emerge as the ghost at the banquet. Think of it as a source of hybrid vigor, a source of unsettling ideas to shake uŒp the established mindset."

Neil Gershenfeld argues that "stealth remedial education for the people running the "Big Five" would be a great output from the book."

Freeman Dyson Freeman, one of the few people alive who knew Wiener, notes that "*The* Human *Use of Human Beings* is one of the best books ever written. Wiener got almost everything right. I will be interested to see what your bunch of wizards will do with it."

The Evolving AI Narrative

Things have changed—and they remain the same. Now AI is everywhere. We have the Internet. We have our smartphones. The founders of the dominant companies—the companies that hold "the whip that lashes us"—have net worths of $65 billion, $90 billion, $130 billion. High-profile individuals such as Elon Musk, Nick Bostrom, Martin Rees, Eliezer Yudkowsky, and the late Stephen Hawking have issued dire warnings about AI, resulting in the ascendancy of well-funded institutes tasked with promoting "Nice AI." But will we, as a species, be able to control a fully realized, unsupervised, self-improving AI? Wiener's warnings and admonitions in *The Human Use of Human Beings* are now very real, and they need to be looked at anew by researchers at the forefront of the AI revolution. Here is Dyson again:

> Wiener became increasingly disenchanted with the "gadget worshipers" whose corporate selfishness brought "motives to automatization that go beyond a legitimate curiosity and are sinful in themselves." He knew the danger was not machines becoming more like humans but humans being treated like machines. "The world of the future will be an ever more demanding struggle against the limitations of our intelligence," he warned in *God & Golem, Inc.*, published in

> 1964, the year of his death, “not a comfortable hammock in which we can lie down to be waited upon by our robot slaves.”

It’s time to examine the evolving AI narrative by identifying the leading members of that mainstream community along with the dissidents, and presenting their counternarratives in their own voices.

The essays that follow thus constitute a much-needed update from the field.

John Brockman
New York, 2019

I met **Seth Lloyd** *in the late 1980s, when new ways of thinking were everywhere: the importance of biological organizing principles, the computational view of mathematics and physical processes, the emphasis on parallel networks, the importance of nonlinear dynamics, the new understanding of chaos, connectionist ideas, neural networks, and parallel distributive processing. The advances in computation during that period provided us with a new way of thinking about knowledge.*

Seth likes to refer to himself as a quantum mechanic. He is internationally known for his work in the field of quantum computation, which attempts to harness the exotic properties of quantum theory, like superposition and entanglement, to solve problems that would take several lifetimes to solve on classical computers.

In the essay that follows, he traces the history of information theory from Norbert Wiener's prophetic insights to the predictions of a technological "singularity" that some would have us believe will supplant the human species. His takeaway on the recent programming method known as deep learning is to call for a more modest set of expectations; he notes that despite AI's enormous advances, robots "still can't tie their own shoes."

It's difficult for me to talk about Seth without referencing his relationship with his friend and professor, the late theoretical physicist Heinz Pagels of Rockefeller University. The graduate student and the professor each had a profound effect on each other's ideas.

In the summer of 1988, I visited Heinz and Seth at the Aspen Center for Physics. Their joint work on the subject of complexity was featured in the current issue of Scientific American*; they were ebullient. That was just two weeks before Heinz's tragic death in a hiking accident while descending Pyramid Peak with Seth. They were talking about quantum computing.*

WRONG, BUT MORE RELEVANT THAN EVER
Seth Lloyd

Seth Lloyd *is a theoretical physicist at MIT, Nam P. Suh Professor in the Department of Mechanical Engineering, and an external professor at the Santa Fe Institute.*

The Human Use of Human Beings, Norbert Wiener's 1950 popularization of his highly influential book *Cybernetics: Control and Communication in the Animal and the Machine* (1948), investigates the interplay between human beings and machines in a world in which machines are becoming ever more computationally capable and powerful. It is a remarkably prescient book, and remarkably wrong. Written at the height of the Cold War, it contains a chilling reminder of the dangers of totalitarian organizations and societies, and of the danger to democracy when it tries to combat totalitarianism with totalitarianism's own weapons.

Wiener's *Cybernetics* looked in close scientific detail at the process of control via feedback. ("Cybernetics," from the ancient Greek for "helmsman," is the etymological basis of our word "governor," which is what James Watt called his pathbreaking feedback control device that transformed the use of steam engines.) Because he was immersed in problems of control, Wiener saw the world as a set of complex, interlocking feedback loops, in which sensors, signals, and actuators such as engines interact via an intricate exchange of signals and information. The engineering applications of *Cybernetics* were tremendously influential and effective, giving rise to rockets, robots, automated assembly lines, and a host of precision-engineering techniques—in other words, to the basis of contemporary industrial society.

Wiener had greater ambitions for cybernetic concepts, however, and in *The Human Use of Human Beings* he spells out his thoughts on its application to topics as diverse as Maxwell's Demon, human language, the brain, insect metabolism, the legal system, the role of technological innovation in government, and religion. These broader applications of cybernetics were an almost unequivocal failure. Vigorously hyped from the late 1940s to the early 1960s—to a degree similar to the hype of computer and communication technology that led to the dotcom crash of 2000-2001—cybernetics delivered satellites and telephone switching systems but generated few if any useful developments in social organization and society at large.

Nearly seventy years later, however, *The Human Use of Human Beings* has more to teach us humans than it did the first time around. Perhaps the most remarkable feature of the book is that it introduces a large number of topics concerning human/machine interactions that are still of considerable relevance. Dark in tone, the book makes several predictions about disasters to come in the second half of the 20th century, many of which are almost identical to predictions made today about the second half of the 21st.

For example, Wiener foresaw a moment in the near future of 1950 in which humans would cede control of society to a cybernetic artificial intelligence, which would then proceed to wreak havoc on humankind. The automation of manufacturing, Wiener predicted, would both create large advances in productivity and displace many workers from their jobs—a sequence of events that did indeed come to pass in the ensuing decades. Unless society could find productive occupations for these displaced workers, Wiener warned, revolt would ensue.

But Wiener failed to foresee crucial technological developments. Like pretty much all technologists of the 1950s, he failed to predict the computer revolution. Computers, he thought, would eventually fall in price from hundreds of thousands of (1950s) dollars to tens of thousands; neither he nor his compeers anticipated the tremendous explosion of computer power that would follow the development of the transistor and the integrated circuit. Finally, because of his emphasis on control, Wiener could not foresee a technological world in which innovation and self-organization bubble up from the bottom rather than being imposed from the top.

Focusing on the evils of totalitarianism (political, scientific, and religious), Wiener saw the world in a deeply pessimistic light. His book warned of the catastrophe that awaited us if we didn't mend our ways, fast. The current world of human beings and machines, more than a half century after its publication, is much more complex, richer, and contains a much wider variety of political, social, and scientific systems than he was able to envisage. The warnings of what will happen if we get it wrong, however—for example, control of the entire Internet by a global totalitarian regime—remain as relevant and pressing today as they were in 1950.

What Wiener Got Right

Wiener's most famous mathematical works focused on problems of signal analysis and the effects of noise. During World War II, he developed techniques for aiming anti-aircraft fire by making models that could predict the future trajectory of an airplane by extrapolating from its past behavior. In *Cybernetics* and in *The Human Use of Human Beings*, Wiener notes that this past behavior includes quirks and habits of the human pilot, thus a mechanized device can predict the behavior of humans. Like Alan Turing, whose Turing Test suggested that computing machines could give responses to questions which were indistinguishable from human responses, Wiener was fascinated by the notion of capturing human behavior by mathematical description. In the 1940s, he applied his knowledge of control and feedback loops to neuro-muscular feedback in living systems, and was responsible for bringing Warren McCulloch and Walter Pitts to MIT, where they did their pioneering work on artificial neural networks.

Wiener's central insight was that the world should be understood in terms of information. Complex systems, such as organisms, brains, and human societies, consist of interlocking feedback loops in which signals exchanged between subsystems result in complex but stable behavior. When feedback loops break down, the system goes unstable. He constructed a compelling picture of how complex biological systems function, a picture that is by and large universally accepted today.

Wiener's vision of information as the central quantity in governing the behavior of complex systems was remarkable at the time. Nowadays, when cars and refrigerators are jammed with microprocessors and much of human society revolves around computers and cell phones connected by the Internet, it seems prosaic to emphasize the centrality of information, computation, and communication. In Wiener's time, however, the first digital computers had only just come into existence, and the Internet was not even a twinkle in the technologist's eye.

Wiener's powerful conception of not just engineered complex systems but all complex systems as revolving around cycles of signals and computation led to tremendous contributions to the development of complex human-made systems. The methods he and others developed for the control of missiles, for example, were later put to work in building the Saturn V moon rocket, one of the crowning engineering achievements of the 20th century. In particular, Wiener's applications of cybernetic concepts to the brain and to computerized perception are the direct precursors of today's neural-network-based deep-learning circuits, and of artificial intelligence itself. But current developments in these fields have diverged from his vision, and their future development may well affect the human uses both of human beings and of machines.

What Wiener Got Wrong

It is exactly in the extension of the cybernetic idea to human beings that Wiener's conceptions missed their target. Setting aside his ruminations on language, law, and human society for the moment, look at a humbler but potentially useful innovation that he thought was imminent in 1950. Wiener notes that prosthetic limbs would be much more effective if their wearers could communicate directly with their prosthetics by their own neural signals, receiving information about pressure and position from the limb and directing its subsequent motion. This turned out to be a much harder problem than Wiener envisaged: Seventy years down the road, prosthetic limbs that incorporate neural feedback are still in the very early stages. Wiener's concept was an excellent one—it's just that the problem of interfacing neural signals with mechanical-electrical devices is hard.

More significantly, Wiener (along with pretty much everyone else in 1950) greatly underappreciated the potential of digital computation. As noted, Wiener's mathematical contributions were to the analysis of signals and noise and his analytic methods apply to continuously varying, or analog, signals. Although he participated in the wartime development of digital computation, he never foresaw the exponential explosion of computing power brought on by the introduction and progressive miniaturization of semiconductor circuits. This is hardly Wiener's fault: The transistor hadn't been invented yet, and the vacuum-tube technology of the digital computers he was familiar with was clunky, unreliable, and unscalable to ever larger devices. In an appendix to the 1948 edition of *Cybernetics,* he anticipates chess-playing computers and predicts that they'll be able to look two or three moves ahead. He might have been surprised to learn that within half a century a computer would beat the human world champion at chess.

Technological Overestimation and the Existential Risks of the Singularity

When Wiener wrote his books, a significant example of technological overestimation was about to occur. The 1950s saw the first efforts at developing artificial intelligence, by researchers such as Herbert Simon, John McCarthy, and Marvin Minsky, who began to program computers to perform simple tasks and to construct rudimentary robots. The success of these initial efforts inspired Simon to declare that "machines will be capable, within twenty years, of doing any work a man can do." Such predictions turned out to be spectacularly wrong. As they became more powerful, computers got better and better at playing chess because they could systematically generate and evaluate a vast selection of possible future moves. But the majority of predictions of AI, e.g., robotic maids, turned out to be illusory. When Deep Blue beat Garry Kasparov at chess in 1997, the most powerful room-cleaning robot was a Roomba, which moved around vacuuming at random and squeaked when it got caught under the couch.

Technological prediction is particularly chancy, given that technologies progress by a series of refinements, halted by obstacles and overcome by innovation. Many obstacles and some innovations can be anticipated, but more cannot. In my own work with experimentalists on building quantum computers, I typically find that some of the technological steps I expect to be easy turn out to be impossible, whereas some of the tasks I imagine to be impossible turn out to be easy. You don't know until you try.

In the 1950s, partly inspired by conversations with Wiener, John von Neumann introduced the notion of the "technological singularity." Technologies tend to improve exponentially, doubling in power or sensitivity over some interval of time. (For example, since 1950, computer technologies have been doubling in power roughly every two years, an observation enshrined as Moore's Law.) Von Neumann extrapolated from the observed exponential rate of technological improvement to predict that "technological progress will become incomprehensively rapid and complicated," outstripping human capabilities in the not too distant future. Indeed, if one extrapolates the growth of raw computing power—expressed in terms of bits and bit flips—into the future at its current rate, computers should match human brains sometime in the next two to four decades (depending on how one estimates the information-processing power of human brains).

The failure of the initial overly optimistic predictions of AI dampened talk about the technological singularity for a few decades, but since the 2005 publication of Ray Kurzweil's *The Singularity is Near,* the idea of technological advance leading to superintelligence is back in force. Some believers, Kurzweil included, regard this singularity as an opportunity: Humans can merge their brains with the superintelligence and thereby live forever. Others, such as Stephen Hawking and Elon Musk, worried that this superintelligence would prove to be malign and regarded it as the greatest existing threat to human civilization. Still others, including some of the contributors to the present volume, think such talk is overblown.

Wiener's life work and his failure to predict its consequences are intimately bound up in the idea of an impending technological singularity. His work on neuroscience and his initial support of McCulloch and Pitts adumbrated the startlingly effective deep-learning methods of the present day. Over the past decade, and particularly in the last five years, such deep-learning techniques have finally exhibited what Wiener liked to call *Gestalt*—for example, the ability to recognize that a circle is a circle even if when slanted sideways it looks like an ellipse. His work on control, combined with his work on neuromuscular feedback, was significant for the development of robotics and is the inspiration for neural-based human/machine interfaces. His lapses in technological prediction, however, suggest that we should take the notion of a technological singularity with a grain of salt. The general difficulties of technological prediction and the problems specific to the development of a superintelligence should warn us against overestimating both the power and the efficacy of information processing.

The Arguments for Singularity Skepticism

No exponential increase lasts forever. An atomic explosion grows exponentially, but only until it runs out of fuel. Similarly, the exponential advances in Moore's Law are starting to run into limits imposed by basic physics. The clock speed of computers maxed out at a few gigahertz a decade and a half ago, simply because the chips were starting to melt. The miniaturization of transistors is already running into quantum-mechanical problems due to tunneling and leakage currents. Eventually, the various exponential improvements in memory and processing driven by Moore's Law will grind to a halt. A few more decades, however, will probably be time enough for the raw information-processing power of computers to match that of brains—at least by the crude measures of number of bits and number of bit-flips per second.

Human brains are intricately constructed, the process of millions of years of natural selection. In Wiener's time, our understanding of the architecture of the brain was rudimentary and simplistic. Since then, increasingly sensitive instrumentation and imaging techniques have shown our brains to be far more varied in structure and complex in function than Wiener could have imagined. I recently asked Tomaso Poggio, one of the pioneers of modern neuroscience, whether he was worried that computers, with their rapidly increasing processing power, would soon emulate the functioning of the human brain. "Not a chance," he replied.

The recent advances in deep learning and neuromorphic computation are very good at reproducing a particular aspect of human intelligence focused on the operation of the brain's cortex, where patterns are processed and recognized. These advances have enabled a computer to beat the world champion not just of chess but of Go, an impressive feat, but they're far short of enabling a computerized robot to tidy a room. (In fact, robots with anything approaching human capability in a broad range of flexible movements are still far away—search "robots falling down." Robots are good at making precision welds on assembly lines, but they still can't tie their own shoes.)

Raw information-processing power does not mean sophisticated information-processing power. While computer power has advanced exponentially, the programs by which computers operate have often failed to advance at all. One of the primary responses of software companies to increased processing power is to add "useful" features which often make the software harder to use. Microsoft Word reached its apex in 1995 and has been slowly sinking under the weight of added features ever since. Once Moore's Law starts slowing down, software developers will be confronted with hard choices between efficiency, speed, and functionality.

A major fear of the singulariteers is that as computers become more involved in designing their own software they'll rapidly bootstrap themselves into achieving superhuman computational ability. But the evidence of machine learning points in the opposite direction. As machines become more powerful and capable of learning, they learn more and more as human beings do—from multiple examples, often under the supervision of human and machine teachers. Education is as hard and slow for computers as it is for teenagers. Consequently, systems based on deep learning are becoming more rather than less human. The skills they bring to learning are not "better than" but "complementary to" human learning: Computer learning systems can identify patterns that humans cannot—and vice versa. The world's best chess players are neither computers nor humans but humans working together with computers. Cyberspace is indeed inhabited by harmful programs, but these primarily take the form of malware—viruses notable for their malign mindlessness, not for their superintelligence.

Whither Wiener

Wiener noted that exponential technological progress is a relatively modern phenomenon and not all of it is good. He regarded atomic weapons and the development of missiles with nuclear warheads as a recipe for the suicide of the human species. He compared the headlong exploitation of the planet's resources with the Mad Tea Party of *Alice in Wonderland:* Having laid waste to one local environment, we make progress simply by moving on to lay waste to the next. Wiener's optimism about the development of computers and neuro-mechanical systems was tempered by his pessimism about their exploitation by authoritarian governments, such as the Soviet Union, and the tendency for democracies, such as the United States, to become more authoritarian themselves in confronting the threat of authoritarianism.

What would Wiener think of the current human use of human beings? He would be amazed by the power of computers and the Internet. He would be happy that the early neural nets in which he played a role have spawned powerful deep-learning systems that exhibit the perceptual ability he demanded of them—although he might not be impressed that one of the most prominent examples of such computerized *Gestalt* is the ability to recognize photos of kittens on the World Wide Web. Rather than regarding machine intelligence as a threat, I suspect he would regard it as a phenomenon in its own right, different from and co-evolving with our own human intelligence.

Unsurprised by global warming—the Mad Tea Party of our era—Wiener would applaud the exponential improvement in alternative-energy technologies and would apply his cybernetic expertise to developing the intricate set of feedback loops needed to incorporate such technologies into the coming smart electrical grid. Nonetheless, recognizing that the solution to the problem of climate change is at least as much political as it is technological, he would undoubtedly be pessimistic about our chances of solving this civilization-threatening problem in time. Wiener hated hucksters—political hucksters most of all—but he acknowledged that hucksters would always be with us.

It's easy to forget just how scary Wiener's world was. The United States and the Soviet Union were in a full-out arms race, building hydrogen bombs mounted on nuclear warheads carried by intercontinental ballistic missiles guided by navigation systems to which Wiener himself—to his dismay—had contributed. I was four years old when Wiener died. In 1964, my nursery school class was practicing duck-and-cover under our desks to prepare for a nuclear attack. Given the human use of human beings in his own day, if he could see our current state, Wiener's first response would be to be relieved that we are still alive.

In the 1980s, **Judea Pearl** *introduced a new approach to artificial intelligence called Bayesian networks. This probability-based model of machine reasoning enabled machines to function—in a complex and uncertain world—as "evidence engines," continuously revising their beliefs in light of new evidence.*

Within a few years, Judea's Bayesian networks had completely overshadowed the previous rule-based approaches to artificial intelligence. The advent of deep learning—in which computers, in effect, teach themselves to be smarter by observing tons of data, has given him pause, because this method lacks transparency.

While recognizing the impressive achievements in deep learning by colleagues such as Michael Jordan and Geoffrey Hinton, he feels uncomfortable with this kind of opacity. He set out to understand the theoretical limitations of deep-learning systems and points out that basic barriers exist that will prevent them from achieving a human kind of intelligence, no matter what we do. Leveraging the computational benefits of Bayesian networks, Judea realized that the combination of simple graphical models and data could also be used to represent and infer cause-effect relationships. The significance of this discovery far transcends its roots in artificial intelligence. His latest book explains causal thinking to the general public; you might say it is a primer on how to think even though human.

Judea's principled, mathematical approach to causality is a profound contribution to the realm of ideas. It has already benefited virtually every field of inquiry, especially the data-intensive health and social sciences.

THE LIMITATIONS OF OPAQUE LEARNING MACHINES

Judea Pearl

***Judea Pearl** is a professor of computer science and director of the Cognitive Systems Laboratory at UCLA. His most recent book, co-authored with Dana Mackenzie, is* The Book of Why: The New Science of Cause and Effect.

As a former physicist, I was extremely interested in cybernetics. Though it did not utilize the full power of Turing Machines, it was highly transparent, perhaps because it was founded on classical control theory and information theory. We are losing this transparency now, with the deep-learning style of machine learning. It is fundamentally a curve-fitting exercise that adjusts weights in intermediate layers of a long input-output chain.

I find many users who say that it "works well and we don't know why." Once you unleash it on large data, deep learning has its own dynamics, it does its own repair and its own optimization, and it gives you the right results most of the time. But when it doesn't, you don't have a clue about what went wrong and what should be fixed. In particular, you do not know if the fault is in the program, in the method, or because things have changed in the environment. We should be aiming at a different kind of transparency.

Some argue that transparency is not really needed. We don't understand the neural architecture of the human brain, yet it runs well, so we forgive our meager understanding and use human helpers to great advantage. In the same way, they argue, why not unleash deep-learning systems and create intelligence without understanding how they work? I buy this argument to some extent. I personally don't like opacity, so I won't spend my time on deep learning, but I know that it has a place in the makeup of intelligence. I know that non-transparent systems can do marvelous jobs, and our brain is proof of that marvel.

But this argument has its limitation. The reason we can forgive our meager understanding of how human brains work is because our brains work the same way, and that enables us to communicate with other humans, learn from them, instruct them, and motivate them in our own native language. If our robots will all be as opaque as AlphaGo, we won't be able to hold a meaningful conversation with them, and that would be unfortunate. We will need to retrain them whenever we make a slight change in the task or in the operating environment.

So, rather than experimenting with opaque learning machines, I am trying to understand their theoretical limitations and examine how these limitations can be overcome. I do it in the context of causal-reasoning tasks, which govern much of how scientists think about the world and, at the same time, are rich in intuition and toy examples, so we can monitor the progress in our analysis. In this context, we've discovered that some basic barriers exist, and that unless they are breached we won't get a real human kind of intelligence no matter what we do. I believe that charting these barriers may be no less important than banging our heads against them.

Current machine-learning systems operate almost exclusively in a statistical, or model-blind, mode, which is analogous in many ways to fitting a function to a cloud of data points. Such systems cannot reason about "what if ?" questions and, therefore, cannot serve as the basis for Strong AI—that is, artificial intelligence that emulates human-level reasoning and competence. To achieve human-level intelligence, learning machines need the guidance of a blueprint of reality, a model—similar to a road map that guides us in driving through an unfamiliar city.

To be more specific, current learning machines improve their performance by optimizing parameters for a stream of sensory inputs received from the environment. It is a slow process, analogous to the natural-selection process that drives Darwinian evolution. It explains how species like eagles and snakes have developed superb vision systems over millions of years. It cannot explain, however, the super-evolutionary process that enabled humans to build eyeglasses and telescopes over barely a thousand years. What humans had that other species lacked was a mental representation of their environment—representations that they could manipulate at will to imagine alternative hypothetical environments for planning and learning.

Historians of *Homo sapiens* such as Yuval Noah Harari and Steven Mithen are in general agreement that the decisive ingredient that gave our ancestors the ability to achieve global dominion about forty thousand years ago was their ability to create and store a mental representation of their environment, interrogate that representation, distort it by mental acts of imagination, and finally answer the "What if?" kind of questions. Examples are interventional questions ("What if I do such-and-such?") and retrospective or counterfactual questions ("What if I had acted differently?"). No learning machine in operation today can answer such questions. Moreover, most learning machines do not possess a representation from which the answers to such questions can be derived.

With regard to causal reasoning, we find that you can do very little with any form of model-blind curve fitting, or any statistical inference, no matter how sophisticated the fitting process is. We have also found a theoretical framework for organizing such limitations, which forms a hierarchy.

On the first level, you have statistical reasoning, which can tell you only how seeing one event would change your belief about another. For example, what can a symptom tell you about a disease?

Then you have a second level, which entails the first but not vice versa. It deals with actions. "What will happen if we raise prices?" "What if you make me laugh?" That second level of the hierarchy requires information about interventions which is not available in the first. This information can be encoded in a graphical model, which merely tells us which variable responds to another.

The third level of the hierarchy is the counterfactual. This is the language used by scientists. "What if the object were twice as heavy?" "What if I were to do things differently?" "Was it the aspirin that cured my headache, or the nap I took?" Counterfactuals are at the top level in the sense that they cannot be derived even if we could predict the effects of all actions. They need an extra ingredient, in the form of equations, to tell us how variables respond to changes in other variables.

One of the crowning achievements of causal-inference research has been the algorithmization of both interventions and counterfactuals, the top two layers of the hierarchy. In other words, once we encode our scientific knowledge in a model (which may be qualitative), algorithms exist that examine the model and determine if a given query, be it about an intervention or about a counterfactual, can be estimated from the available data—and, if so, how. This capability has transformed dramatically the way scientists are doing science, especially in such data-intensive sciences as sociology and epidemiology, for which causal models have become a second language. These disciplines view their linguistic transformation as the Causal Revolution. As Harvard social scientist Gary King puts it, "More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history."

As I contemplate the success of machine learning and try to extrapolate it to the future of AI, I ask myself, "Are we aware of the basic limitations that were discovered in the causal-inference arena? Are we prepared to circumvent the theoretical impediments that prevent us from going from one level of the hierarchy to another level?"

I view machine learning as a tool to get us from data to probabilities. But then we still have to make two extra steps to go from probabilities into real understandingnce—two big steps. One is to predict the effect of actions, and the second is counterfactual imagination. We cannot claim to understand reality unless we make the last two steps.

In his insightful book *Foresight and Understanding* (1961), the philosopher Stephen Toulmin identified the transparency-versus-opacity contrast as the key to understanding the ancient rivalry between Greek and Babylonian sciences. According to Toulmin, the Babylonian astronomers were masters of black-box predictions, far surpassing their Greek rivals in accuracy and consistency of celestial observations. Yet Science favored the creative-speculative strategy of the Greek astronomers, which was wild with metaphorical imagery: circular tubes full of fire, small holes through which celestial fire was visible as stars, and hemispherical Earth riding on turtleback. It was this wild modeling strategy, not Babylonian extrapolation, that jolted Eratosthenes (276-194 BC) to perform one of the most creative experiments in the ancient world and calculate the circumference of the Earth. Such an experiment would never have occurred to a Babylonian data-fitter.

Model-blind approaches impose intrinsic limitations on the cognitive tasks that Strong AI can perform. My general conclusion is that human-level AI cannot emerge solely from model-blind learning machines; it requires the symbiotic collaboration of data and models.

Data science is a science only to the extent that it facilitates the interpretation of data—a two-body problem, connecting data to reality. Data alone are hardly a science, no matter how "big" they get and how skillfully they are manipulated. Opaque learning systems may get us to Babylon, but not to Athens.

Computer scientist **Stuart Russell***, along with Elon Musk, Stephen Hawking, Max Tegmark, and numerous others, has insisted that attention be paid to the potential dangers in creating an intelligence on the superhuman (or even the human) level—an AGI, or artificial general intelligence, whose programmed purposes may not necessarily align with our own.*

His early work was on understanding the notion of "bounded optimality" as a formal definition of intelligence that you can work on. He developed the technique of rational meta-reasoning, "which is, roughly speaking, that you do the computations that you expect to improve the quality of your ultimate decision as quickly as possible." He has also worked on the unification of probability theory and first-order logic—resulting in a new and far more effective monitoring system for the Comprehensive Nuclear Test Ban Treaty—and on the problem of decision making over long timescales (his presentations on the latter topic are usually titled, "Life: Play and Win in 20 trillion moves").

He is very concerned with the continuing development of autonomous weapons, such as lethal micro-drones, which are potentially scalable into weapons of mass destruction. He drafted the letter from forty of the world's leading AI researchers to President Obama which resulted in high-level national-security meetings.

His current work centers on the creation of what he calls "provably beneficial" AI. He wants to ensure AI safety by "imbuing systems with explicit uncertainty" about the objectives of their human programmers, an approach that would amount to a fairly radical reordering of current AI research.

Stuart is also on the radar of anyone who has taken a course in computer science in the last twenty-odd years. He is co-author of "the" definitive AI textbook, with an estimated 5-million-plus English-language readers.

THE PURPOSE PUT INTO THE MACHINE

Stuart Russell

***Stuart Russell** is a professor of computer science and Smith-Zadeh Professor in Engineering at UC Berkeley. He is the coauthor (with Peter Norvig) of* Artificial Intelligence: A Modern Approach.

Among the many issues raised in Norbert Wiener's *The Human Use of Human Beings* (1950) that are currently relevant, the most significant to the AI researcher is the possibility that humanity may cede control over its destiny to machines.

Wiener considered the machines of the near future as far too limited to exert global control, imagining instead that machines and machine-like control systems would be wielded by human elites to reduce the great mass of humanity to the status of "cogs and levers and rods." Looking further ahead, he pointed to the difficulty of correctly specifying objectives for highly capable machines, noting

> a few of the simpler and more obvious truths of life, such as that when a djinnee is found in a bottle, it had better be left there; that the fisherman who craves a boon from heaven too many times on behalf of his wife will end up exactly where he started; that if you are given three wishes, you must be very careful what you wish for.

The dangers are clear enough:

> Woe to us if we let [the machine] decide our conduct, unless we have previously examined the laws of its action, and know fully that its conduct will be carried out on principles acceptable to us! On the other hand, the machine like the djinnee, which can learn and can make decisions on the basis of its learning, will in no way be obliged to make such decisions as we should have made, or will be acceptable to us.

Ten years later, after seeing Arthur Samuel's checker-playing program learn to play checkers far better than its creator, Wiener published "Some Moral and Technical Consequences of Automation" in *Science*. In this paper, the message is even clearer:

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere . . . we had better be quite sure that the purpose put into the machine is the purpose which we really desire. . . .

In my view, this is the source of the existential risk from superintelligent AI cited in recent years by such observers as Elon Musk, Bill Gates, Stephen Hawking, and Nick Bostrom.

Putting Purposes Into Machines

The goal of AI research has been to understand the principles underlying intelligent behavior and to build those principles into machines that can then exhibit such behavior. In the 1960s and 1970s, the prevailing theoretical notion of intelligence was the capacity for logical reasoning, including the ability to derive plans of action guaranteed to achieve a specified goal. More recently, a consensus has emerged around the idea of a rational agent that perceives, and acts in order to maximize, its expected utility. Subfields such as logical planning, robotics, and natural-language understanding are special cases of the general paradigm. AI has incorporated probability theory to handle uncertainty, utility theory to define objectives, and statistical learning to allow machines to adapt to new circumstances. These developments have created strong connections to other disciplines that build on similar concepts, including control theory, economics, operations research, and statistics.

In both the logical-planning and rational-agent views of AI, the machine's objective—whether in the form of a goal, a utility function, or a reward function (as in reinforcement learning)—is specified exogenously. In Wiener's words, this is "the purpose put into the machine." Indeed, it has been one of the tenets of the field that AI systems should be general-purpose—i.e., capable of accepting a purpose as input and then achieving it—rather than special-purpose, with their goal implicit in their design. For example, a self-driving car should accept a destination as input instead of having one fixed destination. However, some aspects of the car's "driving purpose" are fixed, such as that it shouldn't hit pedestrians. This is built directly into the car's steering algorithms rather than being explicit: No self-driving car in existence today "knows" that pedestrians prefer not to be run over.

Putting a purpose into a machine which optimizes its behavior according to clearly defined algorithms seems an admirable approach to ensuring that the machine's "conduct will be carried out on principles acceptable to us!" But, as Wiener warns, we need to put in the right purpose. We might call this the King Midas problem: Midas got exactly what he asked for—namely, that everything he touched would turn to gold—but too late he discovered the drawbacks of drinking liquid gold and eating solid gold. The technical term for putting in the right purpose is *value alignment.* When it fails, we may inadvertently imbue machines with objectives counter to our own. Tasked with finding a cure for cancer as fast as possible, an AI system might elect to use the entire human population as guinea pigs for its experiments. Asked to de-acidify the oceans, it might use up all the oxygen in the atmosphere as a side effect. This is a common characteristic of systems that optimize: Variables not included in the objective may be set to extreme values to help optimize that objective.

Unfortunately, neither AI nor other disciplines (economics, statistics, control theory, operations research) built around the optimization of objectives have much to say about how to identify the purposes "we really desire." Instead, they assume that objectives are simply implanted into the machine. AI research, in its present form, studies the ability to achieve objectives, not the design of those objectives.

Steve Omohundro has pointed to a further difficulty, observing that intelligent entities must act to preserve their own existence. This tendency has nothing to do with a self-preservation instinct or any other biological notion; it's just that an entity cannot achieve its objectives if it's dead. According to Omohundro's argument, a superintelligent machine that has an off-switch—which some, including Alan Turing himself, in a 1951 talk on BBC Radio 3, have seen as our potential salvation—will take steps to disable the switch in some way.[1] Thus we may face the prospect of superintelligent machines—their actions by definition unpredictable by us and their imperfectly specified objectives conflicting with our own—whose motivation to preserve their existence in order to achieve those objectives may be insuperable.

---

[1] Omohundro, "The Basic AI Drives," in *Proc. First AGI Conf.,* 171: "Artificial General Intelligence," eds. P. Wang, B. Goertzel, & S. Franklin (IOS press, 2008).

1001 Reasons to Pay No Attention

Objections have been raised to these arguments, primarily by researchers within the AI community. The objections reflect a natural defensive reaction, coupled perhaps with a lack of imagination about what a superintelligent machine could do. None hold water on closer examination. Here are some of the more common ones:

- *Don't worry, we can just switch it off.*[2] This is often the first thing that pops into a layperson's head when considering risks from superintelligent AI—as if a superintelligent entity would never think of that. This is rather like saying that the risk of losing to DeepBlue or AlphaGo is negligible—all one has to do is make the right moves.
- *Human-level or superhuman AI is impossible.*[3] This is an unusual claim for AI researchers to make, given that, from Turing onward, they have been fending off such claims from philosophers and mathematicians. The claim, which is backed by no evidence, appears to concede that if superintelligent AI *were* possible, it *would* be a significant risk. It's as if a bus driver, with all of humanity as passengers, said, "Yes, I am driving toward a cliff—in fact, I'm pressing the pedal to the metal! But trust me, we'll run out of gas before we get there!" The claim represents a foolhardy bet against human ingenuity. We have made such bets before and lost. On September 11, 1933, renowned physicist Ernest Rutherford stated, with utter confidence, "Anyone who expects a source of power from the transformation of these atoms is talking moonshine." On September 12, 1933, Leo Szilard invented the neutron-induced nuclear chain reaction. A few years later he demonstrated such a reaction in his laboratory at Columbia University. As he recalled in a memoir: "We switched everything off and went home. That night, there was very little doubt in my mind that the world was headed for grief."
- *It's too soon to worry about it.* The right time to worry about a potentially serious problem for humanity depends not just on when the problem will occur but also on how much time is needed to devise and implement a solution that avoids the risk. For example, if we were to detect a large asteroid predicted to collide with the Earth in 2067, would we say, "It's too soon to worry"? And if we consider the global catastrophic risks from climate change predicted to occur later in this century, is it too soon to take action to prevent them? On the contrary, it may be too late. The relevant timescale for human-level AI is less predictable, but, like nuclear fission, it might arrive considerably sooner than expected. One variation on this argument is Andrew Ng's statement that it's "like worrying about overpopulation on Mars." This appeals to a convenient analogy: Not only is the

---

[2] AI researcher Jeff Hawkins, for example, writes, "Some intelligent machines will be virtual, meaning they will exist and act solely within computer networks. . . . It is always possible to turn off a computer network, even if painful." https://www.recode.net/2015/3/2/11559576/.

[3] The AI100 report (Peter Stone *et al.*), sponsored by Stanford University, includes the following: "Unlike in the movies, there is no race of superhuman robots on the horizon or probably even possible." https://ai100.stanford.edu/2016-report.

risk easily managed and far in the future, but also it's extremely unlikely that we'd even try to move billions of humans to Mars in the first place. The analogy is a false one, however. We are already devoting huge scientific and technical resources to creating ever-more-capable AI systems. A more apt analogy would be a plan to move the human race to Mars with no consideration for what we might breathe, drink, or eat once we'd arrived.

- *Human-level AI isn't really imminent, in any case.* The AI100 report, for example, assures us, "Contrary to the more fantastic predictions for AI in the popular press, the Study Panel found no cause for concern that AI is an imminent threat to humankind." This argument simply misstates the reasons for concern, which are not predicated on imminence. In his 2014 book, *Superintelligence: Paths, Dangers, Strategies,* Nick Bostrom, for one, writes, "It is no part of the argument in this book that we are on the threshold of a big breakthrough in artificial intelligence, or that we can predict with any precision when such a development might occur."
- *You're just a Luddite*. It's an odd definition of Luddite that includes Turing, Wiener, Minsky, Musk, and Gates, who rank among the most prominent contributors to technological progress in the 20th and 21st centuries.[4] Furthermore, the epithet represents a complete misunderstanding of the nature of the concerns raised and the purpose for raising them. It is as if one were to accuse nuclear engineers of Luddism if they pointed out the need for control of the fission reaction. Some objectors also use the term "anti-AI," which is rather like calling nuclear engineers "anti-physics." The purpose of understanding and preventing the risks of AI is to ensure that we can realize the benefits. Bostrom, for example, writes that success in controlling AI will result in "a civilizational trajectory that leads to a compassionate and jubilant use of humanity's cosmic endowment"—hardly a pessimistic prediction.
- *Any machine intelligent enough to cause trouble will be intelligent enough to have appropriate and altruistic objectives.*[5] (Often, the argument adds the premise that people of greater intelligence tend to have more altruistic objectives, a view that may be related to the self-conception of those making the argument.) This argument is related to Hume's *is-ought* problem and G. E. Moore's naturalistic fallacy, suggesting that somehow the machine, as a result of its intelligence, will simply *perceive* what is right, given its experience of the world. This is implausible; for example, one cannot perceive, in the design of a chessboard and chess pieces, the goal of checkmate; the same chessboard and pieces can be used for suicide chess, or indeed many other games still to be invented. Put another way: Where Bostrom imagines humans driven extinct by a putative robot that turns the planet into a sea of paper clips, we humans see this outcome as tragic,

---

[4] Elon Musk, Stephen Hawking, and others (including, apparently, the author) received the 2015 Luddite of the Year Award from the Information Technology Innovation Foundation: https://itif.org/publications/2016/01/19/artificial-intelligence-alarmists-win-itif%E2%80%99s-annual-luddite-award.

[5] Rodney Brooks, for example, asserts that it's impossible for a program to be "smart enough that it would be able to invent ways to subvert human society to achieve goals set for it by humans, without understanding the ways in which it was causing problems for those same humans." http://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/.

whereas the iron-eating bacterium *Thiobacillus ferrooxidans* is thrilled. Who's to say the bacterium is wrong? The fact that a machine has been given a fixed objective by humans doesn't mean that it will automatically recognize the importance to humans of things that aren't part of the objective. Maximizing the objective may well cause problems for humans, but, by definition, the machine will not recognize those problems as problematic.

- *Intelligence is multidimensional, "so 'smarter than humans' is a meaningless concept."*[6] It is a staple of modern psychology that IQ doesn't do justice to the full range of cognitive skills that humans possess to varying degrees. IQ is indeed a crude measure of human intelligence, but it is utterly meaningless for current AI systems, because their capabilities across different areas are uncorrelated. How do we compare the IQ of Google's search engine, which cannot play chess, with that of DeepBlue, which cannot answer search queries?

  None of this supports the argument that because intelligence is multifaceted, we can ignore the risk from superintelligent machines. If "smarter than humans" is a meaningless concept, then "smarter than gorillas" is also meaningless, and gorillas therefore have nothing to fear from humans; clearly, that argument doesn't hold water. Not only is it logically possible for one entity to be more capable than another across all the relevant dimensions of intelligence, it is also possible for one species to represent an existential threat to another even if the former lacks an appreciation for music and literature.

Solutions

Can we tackle Wiener's warning head-on? Can we design AI systems whose purposes don't conflict with ours, so that we're sure to be happy with how they behave? On the face of it, this seems hopeless, because it will doubtless prove infeasible to write down our purposes correctly or imagine all the counterintuitive ways a superintelligent entity might fulfill them.

If we treat superintelligent AI systems as if they were black boxes from outer space, then indeed we have no hope. Instead, the approach we seem obliged to take, if we are to have any confidence in the outcome, is to define some formal problem F, and design AI systems to be F-solvers, such that no matter how perfectly a system solves F, we're guaranteed to be happy with the solution. If we can work out an appropriate F that has this property, we'll be able to create *provably beneficial* AI.

Here's an example of how *not* to do it: Let a reward be a scalar value provided periodically by a human to the machine, corresponding to how well the machine has behaved during each period, and let F be the problem of maximizing the expected sum of rewards obtained by the machine. The optimal solution to this problem is not, as one might hope, to behave well, but instead to take control of the human and force him or her to provide a stream of maximal rewards. This is known as the wireheading problem, based on observations that humans themselves are susceptible to the same problem if given a means to electronically stimulate their own pleasure centers.

There is, I believe, an approach that may work. Humans can reasonably be described as having (mostly implicit) preferences over their future lives—that is, given enough time and unlimited visual aids, a human could express a preference (or indifference) when offered a choice between two future lives laid out before him or her in all their aspects. (This idealization ignores the possibility that our minds are composed of subsystems with incompatible preferences; if true, that would limit a machine's ability to optimally satisfy our preferences, but it doesn't seem to prevent us from designing machines that avoid catastrophic outcomes.) The formal problem F to be solved by the machine in this case is to maximize human future-life preferences subject to its initial uncertainty as to what they are. Furthermore, although the future-life preferences are hidden variables, they're grounded in a voluminous source of evidence—namely, all of the human choices ever made. This formulation sidesteps Wiener's problem: The machine may learn more about human preferences as it goes along, of course, but it will never achieve complete certainty.

---

[6] Kevin Kelly, "The Myth of a Superhuman AI," *Wired*, Apr. 25, 2017.

A more precise definition is given by the framework of cooperative inverse-reinforcement learning, or CIRL. A CIRL problem involves two agents, one human and the other a robot. Because there are two agents, the problem is what economists call a game. It is a game of partial information, because while the human knows the reward function, the robot doesn't—even though the robot's job is to maximize it.

A simple example: Suppose that Harriet, the human, likes to collect paper clips and staples and her reward function depends on how many of each she has. More precisely, if she has p paper clips and s staples, her degree of happiness is $\theta p + (1-\theta)s$, where θ *is* essentially an exchange rate between paper clips and staples. If θ is 1, she likes only paper clips; if θ is 0, she likes only staples; if θ is 0.5, she is indifferent between them; and so on. It's the job of Robby, the robot, to produce the paper clips and staples. The point of the game is that Robby wants to make Harriet happy, but he doesn't know the value of θ, so he isn't sure how many of each to produce.

Here's how the game works. Let the true value of θ be 0.49—that is, Harriet has a slight preference for staples over paper clips. And let's assume that Robby has a uniform prior belief about θ—that is, he believes θ is equally likely to be any value between 0 and 1. Harriet now gets to do a small demonstration, producing either two paper clips or two staples or one of each. After that, the robot can produce either ninety paper clips, or ninety staples, or fifty of each. You might think that Harriet, who prefers staples to paper clips, should produce two staples. But in that case, Robby's rational response would be to produce ninety staples (with a total value to Harriet of 45.9), which is a less desirable outcome for Harriet than fifty of each (total value 50.0). The optimal solution of this particular game is that Harriet produces one of each, so then Robby makes fifty of each. Thus, the way the game is defined encourages Harriet to "teach" Robby—as long as she knows that Robby is watching carefully.

Within the CIRL framework, one can formulate and solve the off-switch problem—that is, the problem of how to prevent a robot from disabling its off-switch. (Turing may rest easier.) A robot that's uncertain about human preferences actually benefits from being switched off, because it understands that the human will press the off-switch to prevent the robot from doing something counter to those preferences. Thus the robot is incentivized to preserve the off-switch, and this incentive derives directly from its uncertainty about human preferences.[7]

The off-switch example suggests some templates for controllable-agent designs and provides at least one case of a provably beneficial system in the sense introduced above. The overall approach resembles mechanism-design problems in economics, wherein one incentivizes other agents to behave in ways beneficial to the designer. The key difference here is that we are building one of the agents in order to benefit the other.

---

[7] See Hadfield-Menell et al., "The Off-Switch Game," https://arxiv.org/pdf/1611.08219.pdf.

There are reasons to think this approach may work in practice. First, there is abundant written and filmed information about humans doing things (and other humans reacting). Technology to build models of human preferences from this storehouse will presumably be available long before superintelligent AI systems are created. Second, there are strong, near-term economic incentives for robots to understand human preferences: If one poorly designed domestic robot cooks the cat for dinner, not realizing that its sentimental value outweighs its nutritional value, the domestic-robot industry will be out of business.

There are obvious difficulties, however, with an approach that expects a robot to learn underlying preferences from human behavior. Humans are irrational, inconsistent, weak-willed, and computationally limited, so their actions don't always reflect their true preferences. (Consider, for example, two humans playing chess. Usually, one of them loses, but not on purpose!) So robots can learn from nonrational human behavior only with the aid of much better cognitive models of humans. Furthermore, practical and social constraints will prevent all preferences from being maximally satisfied simultaneously, which means that robots must mediate among conflicting preferences—something that philosophers and social scientists have struggled with for millennia. And what should robots learn from humans who enjoy the suffering of others? It may be best to zero out such preferences in the robots' calculations.

Finding a solution to the AI control problem is an important task; it may be, in Bostrom's words, "the essential task of our age." Up to now, AI research has focused on systems that are better at making decisions, but this is not the same as making better decisions. No matter how excellently an algorithm maximizes, and no matter how accurate its model of the world, a machine's decisions may be ineffably stupid in the eyes of an ordinary human if its utility function is not well aligned with human values.

This problem requires a change in the definition of AI itself—from a field concerned with pure intelligence, independent of the objective, to a field concerned with systems that are provably beneficial for humans. Taking the problem seriously seems likely to yield new ways of thinking about AI, its purpose, and our relationship to it.

In 2005, **George Dyson***, a historian of science and technology, visited Google at the invitation of some Google engineers. The occasion was the sixtieth anniversary of John von Neumann's proposal for a digital computer. After the visit, George wrote an essay, "Turing's Cathedral," which, for the first time, alerted the public about what Google's founders had in store for the world. "We are not scanning all those books to be read by people," explained one of his hosts after his talk. "We are scanning them to be read by an AI."*

George offers a counternarrative to the digital age. His interests have included the development of the Aleut kayak, the evolution of digital computing and telecommunications, the origins of the digital universe, and a path not taken into space. His career (he never finished high school, yet has been awarded an honorary doctorate from the University of Victoria) has proved as impossible to classify as his books.

He likes to point out that analog computing, once believed to be as extinct as the differential analyzer, has returned. He argues that while we may use digital components, at a certain point the analog computing being performed by the system far exceeds the complexity of the digital code with which it is built. He believes that true artificial intelligence—with analog control systems emerging from a digital substrate the way digital computers emerged out of analog components in the aftermath of World War II—may not be as far off as we think.

In this essay, George contemplates the distinction between analog and digital computation and finds analog to be alive and well. Nature's response to an attempt to program machines to control everything may be machines without programming over which no one has control.

THE THIRD LAW
George Dyson

***George Dyson** is a historian of science and technology and the author of* Baidarka: the Kayak, Darwin Among the Machines, Project Orion, *and* Turing's Cathedral.

The history of computing can be divided into an Old Testament and a New Testament: before and after electronic digital computers and the codes they spawned proliferated across the Earth. The Old Testament prophets, who delivered the underlying logic, included Thomas Hobbes and Gottfried Wilhelm Leibniz. The New Testament prophets included Alan Turing, John von Neumann, Claude Shannon, and Norbert Wiener. They delivered the machines.

Alan Turing wondered what it would take for machines to become intelligent. John von Neumann wondered what it would take for machines to self-reproduce. Claude Shannon wondered what it would take for machines to communicate reliably, no matter how much noise intervened. Norbert Wiener wondered how long it would take for machines to assume control.

Wiener's warnings about control systems beyond human control appeared in 1949, just as the first generation of stored-program electronic digital computers were introduced. These systems required direct supervision by human programmers, undermining his concerns. What's the problem, as long as programmers are in control of the machines? Ever since, debate over the risks of autonomous control has remained associated with the debate over the powers and limitations of digitally coded machines. Despite their astonishing powers, little real autonomy has been observed. This is a dangerous assumption. What if digital computing is being superseded by something else?

Electronics underwent two fundamental transitions over the past hundred years: from analog to digital and from vacuum tubes to solid state. That these transitions occurred together does not mean they are inextricably linked. Just as digital computation was implemented using vacuum tube components, analog computation can be implemented in solid state. Analog computation is alive and well, even though vacuum tubes are commercially extinct.

There is no precise distinction between analog and digital computing. In general, digital computing deals with integers, binary sequences, deterministic logic, and time that is idealized into discrete increments, whereas analog computing deals with real numbers, nondeterministic logic, and continuous functions, including time as it exists as a continuum in the real world.

Imagine you need to find the middle of a road. You can measure its width using any available increment and then digitally compute the middle to the nearest increment. Or you can use a piece of string as an analog computer, mapping the width of the road to the length of the string and finding the middle, without being limited to increments, by doubling the string back upon itself.

Many systems operate across both analog and digital regimes. A tree integrates a wide range of inputs as continuous functions, but if you cut down that tree, you find that it has been counting the years digitally all along.

In analog computing, complexity resides in network topology, not in code. Information is processed as continuous functions of values such as voltage and relative pulse frequency rather than by logical operations on discrete strings of bits. Digital computing, intolerant of error or ambiguity, depends upon error correction at every step along the way. Analog computing tolerates errors, allowing you to live with them.

Nature uses digital coding for the storage, replication, and recombination of sequences of nucleotides, but relies on analog computing, running on nervous systems, for intelligence and control. The genetic system in every living cell is a stored-program computer. Brains aren't.

Digital computers execute transformations between two species of bits: bits representing differences in space and bits representing differences in time. The transformations between these two forms of information, sequence and structure, are governed by the computer's programming, and as long as computers require human programmers, we retain control.

Analog computers also mediate transformations between two forms of information: structure in space and behavior in time. There is no code and no programming. Somehow—and we don't fully understand how—Nature evolved analog computers known as nervous systems, which embody information absorbed from the world. They learn. One of the things they learn is control. They learn to control their own behavior, and they learn to control their environment to the extent that they can.

Computer science has a long history—going back to before there even *was* computer science—of implementing neural networks, but for the most part these have been simulations of neural networks by digital computers, not neural networks as evolved in the wild by Nature herself. This is starting to change: from the bottom up, as the threefold drivers of drone warfare, autonomous vehicles, and cell phones push the development of neuromorphic microprocessors that implement actual neural networks, rather than simulations of neural networks, directly in silicon (and other potential substrates); and from the top down, as our largest and most successful enterprises increasingly turn to analog computation in their infiltration and control of the world.

While we argue about the intelligence of digital computers, analog computing is quietly supervening upon the digital, in the same way that analog components like vacuum tubes were repurposed to build digital computers in the aftermath of World War II. Individually deterministic finite-state processors, running finite codes, are forming large-scale, nondeterministic, non-finite-state metazoan organisms running wild in the real world. The resulting hybrid analog/digital systems treat streams of bits collectively, the way the flow of electrons is treated in a vacuum tube, rather than individually, as bits are treated by the discrete-state devices generating the flow. Bits are the new electrons. Analog is back, and its nature is to assume control.

Governing everything from the flow of goods to the flow of traffic to the flow of ideas, these systems operate statistically, as pulse-frequency coded information is processed in a neuron or a brain. The emergence of intelligence gets the attention of *Homo sapiens,* but what we should be worried about is the emergence of control.

~ ~ ~

Imagine it is 1958 and you are trying to defend the continental United States against airborne attack. To distinguish hostile aircraft, one of the things you need, besides a network of computers and early-warning radar sites, is a map of all commercial air traffic, updated in real time. The United States built such a system and named it SAGE (Semi-Automatic Ground Environment). SAGE in turn spawned Sabre, the first integrated reservation system for booking airline travel in real time. Sabre and its progeny soon became not just a map as to what seats were available but also a system that began to control, with decentralized intelligence, where airliners would fly, and when.

But isn't there a control room somewhere, with someone at the controls? Maybe not. Say, for example, you build a system to map highway traffic in real time, simply by giving cars access to the map in exchange for reporting their own speed and location at the time. The result is a fully decentralized control system. Nowhere is there any controlling model of the system except the system itself.

Imagine it is the first decade of the 21st century and you want to track the complexity of human relationships in real time. For social life at a small college, you could construct a central database and keep it up to date, but its upkeep would become overwhelming if taken to any larger scale. Better to pass out free copies of a simple semi-autonomous code, hosted locally, and let the social network update itself. This code is executed by digital computers, but the analog computing performed by the system as a whole far exceeds the complexity of the underlying code. The resulting pulse-frequency coded model of the social graph *becomes* the social graph. It spreads wildly across the campus and then the world.

What if you wanted to build a machine to capture what everything known to the human species *means*? With Moore's Law behind you, it doesn't take too long to digitize all the information in the world. You scan every book ever printed, collect every email ever written, and gather forty-nine years of video every twenty-four hours, while tracking where people are and what they do, in real time. But how do you capture the *meaning*?

Even in the age of all things digital, this cannot be defined in any strictly logical sense, because meaning, among humans, isn't fundamentally logical. The best you can do, once you have collected all possible answers, is to invite well-defined questions and compile a pulse-frequency weighted map of how everything connects. Before you know it, your system will not only be observing and mapping the meaning of things, it will start *constructing* meaning as well. In time, it will *control* meaning, in the same way as the traffic map starts to control the flow of traffic even though no one seems to be in control.

~ ~ ~

There are three laws of artificial intelligence. The first, known as Ashby's Law, after cybernetician W. Ross Ashby, author of *Design for a Brain*, states that any effective control system must be as complex as the system it controls.

The second law, articulated by John von Neumann, states that the defining characteristic of a complex system is that it constitutes its own simplest behavioral description. The simplest complete model of an organism is the organism itself. Trying to reduce the system's behavior to any formal description makes things more complicated, not less.

The third law states that any system simple enough to be understandable will not be complicated enough to behave intelligently, while any system complicated enough to behave intelligently will be too complicated to understand.

The Third Law offers comfort to those who believe that until we understand intelligence, we need not worry about superhuman intelligence arising among machines. But there is a loophole in the Third Law. It is entirely possible to build something without understanding it. You don't need to fully understand how a brain works in order to build one that works. This is a loophole that no amount of supervision over algorithms by programmers and their ethical advisors can ever close. Provably "good" AI is a myth. Our relationship with true AI will always be a matter of faith, not proof.

We worry too much about machine intelligence and not enough about self-reproduction, communication, and control. The next revolution in computing will be signaled by the rise of analog systems over which digital programming no longer has control. Nature's response to those who believe they can build machines to control everything will be to allow them to build a machine that controls them instead.

Dan Dennett *is the philosopher of choice in the AI community. He is perhaps best known in cognitive science for his concept of intentional systems and his model of human consciousness, which sketches a computational architecture for realizing the stream of consciousness in the massively parallel cerebral cortex. That uncompromising computationalism has been opposed by philosophers such as John Searle, David Chalmers, and the late Jerry Fodor, who have protested that the most important aspects of consciousness—intentionality and subjective qualia—cannot be computed.*

Twenty-five years ago, I was visiting Marvin Minsky, one of the original AI pioneers, and asked him about Dan. "He's our best current philosopher—the next Bertrand Russell," said Marvin, adding that unlike traditional philosophers, Dan was a student of neuroscience, linguistics, artificial intelligence, computer science, and psychology: "He's redefining and reforming the role of the philosopher. Of course, Dan doesn't understand my Society-of-Mind theory, but nobody's perfect."

Dan's view of the efforts of AI researchers to create superintelligent AIs is relentlessly levelheaded. What, me worry? In this essay, he reminds us that AIs, above all, should be regarded—and treated—as tools and not as humanoid colleagues.

He has been interested in information theory since his graduate school days at Oxford. In fact, he told me that early in his career he was keenly interested in writing a book about Wiener's cybernetic ideas. As a thinker who embraces the scientific method, one of his charms is his willingness to be wrong. Of a recent piece entitled "What Is Information?" he has announced, "I stand by it, but it's under revision. I'm already moving beyond it and realizing there's a better way of tackling some of these issues." He will most likely remain cool and collected on the subject of AI research, although he has acknowledged, often, that his own ideas evolve—as anyone's ideas should.

WHAT CAN WE DO?

Daniel C. Dennett

***Daniel C. Dennett** is University Professor and Austin B. Fletcher Professor of Philosophy and director of the Center for Cognitive Studies at Tufts University. He is the author of a dozen books, including* Consciousness Explained *and, most recently,* From Bacteria to Bach and Back: The Evolution of Minds.

Many have reflected on the irony of reading a great book when you are too young to appreciate it. Consigning a classic to the *already read* stack and thereby insulating yourself against any further influence while gleaning only a few ill-understood ideas from it is a recipe for neglect that is seldom benign. This struck me with particular force when I reread *The Human Use of Human Beings* more than sixty years after my juvenile encounter. We should all make it a regular practice to reread books from our youth, where we are apt to discover clear previews of some of our own later “discoveries” and “inventions,” along with a wealth of insights to which we were bound to be impervious until our minds had been torn and tattered, exercised and enlarged by confrontations with life’s problems.

Writing at a time when vacuum tubes were still the primary electronic building blocks and there were only a few actual computers in operation, Norbert Wiener imagined the future we now contend with in impressive detail and with few clear mistakes. Alan Turing’s famous 1950 article “Computing Machinery and Intelligence,” in the philosophy journal *Mind,* foresaw the development of AI, and so did Wiener, but Wiener saw farther and deeper, recognizing that AI would not just imitate—and replace—human beings in many intelligent activities but change human beings in the process.

> We are but whirlpools in a river of ever-flowing water. We are not stuff that abides, but patterns that perpetuate themselves. (p. 96)

When that was written, it could be comfortably dismissed as yet another bit of Heraclitean overstatement. Yeah, yeah, you can never step in the same river twice. But it contains the seeds of the revolution in outlook. Today we know how to think about complex adaptive systems, strange attractors, extended minds, and homeostasis, a change in perspective that promises to erase the “explanatory gap”[8] between mind and mechanism, spirit and matter, a gap that is still ardently defended by latter-day Cartesians who cannot bear the thought that we—we *ourselves*—are self-perpetuating patterns of information-bearing matter, not “stuff that abides.” Those patterns are remarkably resilient and self-restoring but at the same time protean, opportunistic, selfish exploiters of whatever new is available to harness in their quest for perpetuation. And here is where things get dicey, as Wiener recognized. When attractive opportunities abound, we are apt to be willing to pay a little and accept some small, even trivial, cost-of-doing-business for access to new powers. And pretty soon we become so dependent on our new tools that we lose the ability to thrive without them. Options become obligatory.

---

[8] Joseph Levine, “Materialism and Qualia: The Explanatory Gap,” *Pacific Philosophical Quarterly* 64, pp. 354-61 (1983).

It's an old, old story, with many well-known chapters in evolutionary history. Most mammals can synthesize their own vitamin C, but primates, having opted for a diet composed largely of fruit, lost the innate ability. We are now obligate ingesters of vitamin C, but not obligate frugivores like our primate cousins, since we have opted for technology that allows us to make, and take, vitamins as needed. The self-perpetuating patterns that we call human beings are now dependent on clothes, cooked food, vitamins, vaccinations, . . . credit cards, smartphones, and the Internet. And—tomorrow if not already today—AI.

Wiener foresaw the problems that Turing and the other optimists have largely overlooked. The real danger, he said, is

> that such machines, though helpless by themselves, may be used by a human being or a block of human beings to increase their control over the rest of the race or that political leaders may attempt to control their populations by means not of machines themselves but through political techniques as narrow and indifferent to human possibility as if they had, in fact, been conceived mechanically. (p. 181)

The power, he recognized, lay primarily in the algorithms, not the hardware they run on, although the hardware of today makes practically possible algorithms that would have seemed preposterously cumbersome in Wiener's day. What can we say about these "techniques" that are "narrow and indifferent to human possibility"? They have been introduced again and again, some obviously benign, some obviously dangerous, and many in the omnipresent middle ground of controversy.

Consider a few of the skirmishes. My late friend Joe Weizenbaum, Wiener's successor as MIT's Jeremiah of hi-tech, loved to observe that credit cards, whatever their virtues, also provided an inexpensive and almost foolproof way for the government, or corporations, to track the travels and habits and desires of individuals. The anonymity of cash has been largely underappreciated, except by drug dealers and other criminals, and now it may be going extinct. This may make money laundering a more difficult technical challenge in the future, but the AI pattern finders arrayed against it have the side effect of making us all more transparent to any "block of human beings" that may "attempt to control" us.

Looking to the arts, the innovation of digital audio and video recording lets us pay a small price (in the eyes of all but the most ardent audiophiles and film lovers) when we abandon analog formats, and in return provides easy—all too easy?—reproduction of artworks with almost perfect fidelity. But there is a huge hidden cost. Orwell's Ministry of Truth is now a practical possibility. AI techniques for creating all-but-undetectable forgeries of "recordings" of encounters are now becoming available which will render obsolete the tools of investigation we have come to take for granted in the last hundred and fifty years. Will we simply abandon the brief Age of Photographic Evidence and return to the earlier world in which human memory and trust provided the gold standard, or will we develop new techniques of defense and offense in the arms race of truth? (We can imagine a return to *analog* film-exposed-to-light, kept in "tamper-proof" systems until shown to juries, etc., but how long would it be before somebody figured out a way to infect such systems with doubt? One of the disturbing lessons of recent experience is that the task of destroying a reputation for credibility is much less expensive than the task of protecting such a reputation.) Wiener saw the phenomenon at its most general: "…in the long run, there is no distinction between arming ourselves and arming our enemies." (p. 129) The Information Age is also the Dysinformation Age.

What can we do? We need to rethink our priorities with the help of the passionate but flawed analyses of Wiener, Weizenbaum, and the other serious critics of our technophilia. A key phrase, it seems to me, is Wiener's almost offhand observation, above, that "these machines" are "helpless by themselves." As I have been arguing recently, we're making tools, not colleagues, and the great danger is not appreciating the difference, which we should strive to accentuate, marking and defending it with political and legal innovations.

Perhaps the best way to see what is being missed is to note that Alan Turing himself suffered an entirely understandable failure of imagination in his formulation of the famous Turing Test. As everyone knows, it is an adaptation of his "imitation game," in which a man, hidden from view and communicating verbally with a judge, tries to convince the judge that he is in fact a woman, while a woman, also hidden and communicating with the judge, tries to convince the judge that she is the woman. Turing reasoned that this would be a demanding challenge for a man (or for a woman pretending to be a man), exploiting a wealth of knowledge about how the other sex thinks and acts, what they tend to favor or ignore. Surely (*ding!*)[9], any man who could beat a woman at being perceived to be a woman would be an intelligent agent. What Turing did not foresee is the power of deep-learning AI to acquire this wealth of information in an exploitable form *without having to understand it.* Turing imagined an astute and imaginative (and hence conscious) agent who cunningly designed his responses based on his detailed "theory" of what women are likely to do and say. Top-down intelligent design, in short. He certainly didn't think that a man, winning the imitation game, would somehow *become* a woman; he imagined that there would still be a man's consciousness guiding the show. The hidden premise in Turing's almost-argument was: Only a conscious, intelligent *agent* could devise and control a winning strategy in the imitation game. And so it was persuasive to Turing (and others, including me, still a stalwart defender of the Turing Test) to argue that a "computing machine" that could pass as human in a contest with a human might not be conscious in just the way a human being is, but would nevertheless have to be a conscious agent of *some* kind. I think this is still a defensible position—the only defensible position—but you have to understand how resourceful and ingenious a judge would have to be to expose the shallowness of the façade that a deep-learning AI (a tool, not a colleague) could present.

What Turing didn't foresee is the uncanny ability of superfast computers to sift mindlessly through Big Data, of which the Internet provides an inexhaustible supply, finding probabilistic patterns in human activity that could be used to pop "authentic"-seeming responses into the output for almost any probe a judge would think to offer. Wiener also underestimates this possibility, seeing the tell-tale weakness of a machine in not being able to

> take into account the vast range of probability that characterizes the human situation. (p.181)

---

[9] The *surely alarm* (the habit of having a bell ring in your head whenever you see the word in an argument) is described and defended by me in *Intuition Pumps and Other Tools for Thinking* (2013).

But taking into account that range of probability is just where the new AI excels. The only chink in the armor of AI is that word "vast"; human possibilities, thanks to language and the culture that it spawns, are truly Vast.[10] No matter how many patterns we may find with AI in the flood of data that has so far found its way onto the Internet, there are Vastly more possibilities that have never been recorded there. Only a fraction (but not a Vanishing fraction) of the world's accumulated wisdom and design and repartee and silliness has made it onto the Internet, but probably a better tactic for the judge to adopt when confronting a candidate in the Turing Test is not to *search* for such items but to *create* them anew. AI in its current manifestations is parasitic on human intelligence. It quite indiscriminately gorges on whatever has been produced by human creators and extracts the patterns to be found there—including some of our most pernicious habits.[11] These machines do not (yet) have the goals or strategies or capacities for self-criticism and innovation to permit them to transcend their databases by reflectively thinking about their own thinking and their own goals. They are, as Wiener says, helpless, not in the sense of being *shackled* agents or *disabled* agents but in the sense of not being agents at all—not having the capacity to be "moved by reasons" (as Kant put it) presented to them. It is important that we keep it that way, which will take some doing.

One of the flaws in Weizenbaum's book *Computer Power and Human Reason,* something I tried in vain to convince him of in many hours of discussion, is that he could never decide which of two theses he wanted to defend: *AI is impossible!* or *AI is possible but evil!* He wanted to argue, with John Searle and Roger Penrose, that "Strong AI" is impossible, but there are no good arguments for that conclusion. After all, everything we now know suggests that, as I have put it, we are robots made of robots made of robots. . . down to the motor proteins and their ilk, with no magical ingredients thrown in along the way. Weizenbaum's more important and defensible message was that we should not strive to create Strong AI and should be extremely cautious about the AI systems that we can create and have already created. As one might expect, the defensible thesis is a hybrid: *AI* (Strong AI) *is possible in principle but not desirable. The AI that's practically possible is not necessarily evil—unless it is mistaken for Strong AI!*

The gap between today's systems and the science-fictional systems dominating the popular imagination is still huge, though many folks, both lay and expert, manage to underestimate it. Let's consider IBM's Watson, which can stand as a worthy landmark for our imaginations for the time being. It is the result of a very large-scale R&D process extending over many person-centuries of intelligent design, and as George Church notes in these pages, it uses thousands of times more energy than a human brain (a technological limitation that, as he also notes, may be temporary). Its victory in *Jeopardy!* was a genuine triumph, made possible by the formulaic restrictions of the *Jeopardy!* rules, but in order for it to compete, even these rules had to be revised (one of those trade-offs: you give up a little versatility, a little humanity, and get a crowd-pleasing show). Watson is not good company, in spite of misleading ads from IBM that suggest a general conversational ability, and turning Watson into a plausibly multidimensional *agent* would be like turning a hand calculator into Watson. Watson could be a useful core faculty for such an agent, but more like a cerebellum or an amygdala than a mind—at best, a special-purpose subsystem that could play a big supporting role, but not remotely up to the task of framing purposes and plans and building insightfully on its conversational experiences.

---

[10] In *Darwin's Dangerous Idea,* 1995, p. 109, I coined the capitalized version, *Vast,* meaning *Very much more than ASTronomical*, and its complement, *Vanishing,* to replace the usual exaggerations *infinite* and *infinitesimal* for discussions of those possibilities that are not officially infinite but nevertheless infinite for all practical purposes.

[11] Aylin Caliskan-Islam, Joanna J. Bryson & Arvind Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science,* 14 April 2017, 356: 6334, pp. 183-6. DOI: 10.1126/science.aal4230.

Why would we want to create a thinking, creative agent out of Watson? Perhaps Turing's brilliant idea of an operational test has lured us into a trap: the quest to create at least the illusion of a real person behind the screen, bridging the "uncanny valley." The danger, here, is that ever since Turing posed his challenge—which was, after all, a challenge to *fool* the judges—AI creators have attempted to paper over the valley with cutesy humanoid touches, Disneyfication effects that will enchant and disarm the uninitiated. Weizenbaum's ELIZA was the pioneer example of such superficial illusion-making, and it was his dismay at the ease with which his laughably simple and shallow program could persuade people they were having a serious heart-to-heart conversation that first sent him on his mission.

He was right to be worried. If there is one thing we have learned from the restricted Turing Test competitions for the Loebner Prize, it is that even very intelligent people who aren't tuned in to the possibilities and shortcuts of computer programming are readily taken in by simple tricks. The attitudes of people in AI toward these methods of dissembling at the "user interface" have ranged from contempt to celebration, with a general appreciation that the tricks are not deep but can be potent. One shift in attitude that would be very welcome is a candid acknowledgment that humanoid embellishments are *false advertising*—something to condemn, not applaud.

How could that be accomplished? Once we recognize that people are starting to make life-or-death decisions largely on the basis of "advice" from AI systems whose inner operations are unfathomable in practice, we can see a good reason why those who in any way encourage people to put more trust in these systems than they warrant should be held morally and legally accountable. AI systems are very powerful tools—so powerful that even experts will have good reason not to trust their own judgment over the "judgments" delivered by their tools. But then, if these tool users are going to benefit, financially or otherwise, from driving these tools through *terra incognita,* they need to make sure they know how to do this responsibly, with maximum control and justification. Licensing and bonding operators, just as we license pharmacists (and crane operators!) and other specialists whose errors and misjudgments can have dire consequences, can, with pressure from insurance companies and other underwriters, oblige creators of AI systems to go to extraordinary lengths to search for and reveal weaknesses and gaps in their products, and to train those entitled to operate them.

One can imagine a sort of inverted Turing Test in which the judge is on trial; until he or she can spot the weaknesses, the overstepped boundaries, the gaps in a system, no license to operate will be issued. The mental training required to achieve certification as a judge will be demanding. The urge to adopt the intentional stance, our normal tactic whenever we encounter what seems to be an intelligent agent, is almost overpoweringly strong. Indeed, the capacity to resist the allure of treating an apparent person as a person is an ugly talent, reeking of racism or species-ism. Many people would find the cultivation of such a ruthlessly skeptical approach morally repugnant, and we can anticipate that even the most proficient system-users would occasionally succumb to the temptation to "befriend" their tools, if only to assuage their discomfort with the execution of their duties. No matter how scrupulously the AI designers launder the phony "human" touches out of their wares, we can expect novel habits of thought, conversational gambits and ruses, traps and bluffs to arise in this novel setting for human action. The comically long lists of known side effects of new drugs advertised on television will be dwarfed by the obligatory revelations of the sorts of questions that can*not* be responsibly answered by particular systems, with heavy penalties for those who "overlook" flaws in their products. It is widely noted that a considerable part of the growing economic inequality in today's world is due to the wealth accumulated by digital entrepreneurs; we should enact legislation that puts their deep pockets in escrow for the public good. Some of the deepest pockets are voluntarily out in front of these obligations to serve society first and make money secondarily, but we shouldn't rely on good will alone.

We don't need artificial conscious agents. There is a surfeit of natural conscious agents, enough to handle whatever tasks should be reserved for such special and privileged entities. We need intelligent tools. Tools do not have rights, and should not have feelings that could be hurt, or be able to respond with resentment to "abuses" rained on them by inept users.[12] One of the reasons for not making artificial conscious agents is that however autonomous they might become (and in principle, they can be as autonomous, as self-enhancing or self-creating, as any person), they would not—without special provision, which might be waived—share with us natural conscious agents our vulnerability or our mortality.

I once posed a challenge to students in a seminar at Tufts I co-taught with Matthias Scheutz on artificial agents and autonomy: Give me the specs for a robot that could sign a binding contract with you—not as a surrogate for some human owner but on its own. This isn't a question of getting it to understand the clauses or manipulate a pen on a piece of paper but of having and *deserving* legal status as a morally responsible agent. Small children can't sign such contracts, nor can those disabled people whose legal status requires them to be under the care and responsibility of guardians of one sort or another. The problem for robots who might want to attain such an exalted status is that, like Superman, they are too invulnerable to be able to make a credible promise. If they were to renege, what would happen? What would be the penalty for promise-breaking? Being locked in a cell or, more plausibly, dismantled? Being locked up is barely an inconvenience for an AI unless we first install artificial wanderlust that cannot be ignored or disabled by the AI on its own (and it would be systematically difficult to make this a foolproof solution, given the presumed cunning and self-knowledge of the AI); and dismantling an AI (either a robot or a bedridden agent like Watson) is not killing it, if the information stored in its design and software is preserved. The very ease of digital recording and transmitting—the breakthrough that permits software and data to be, in effect, immortal—removes robots from the world of the vulnerable (at least robots of the usually imagined sorts, with digital software and memories). If this isn't obvious, think about how human morality would be affected if we could make "backups" of people every week, say. Diving headfirst on Saturday off a high bridge without benefit of bungee cord would be a rush that you wouldn't remember when your Friday night backup was put online Sunday morning, but you could enjoy the videotape of your apparent demise thereafter.

---

[12] Joanna J. Bryson, "Robots Should Be Slaves," in *Close Engagement with Artificial Companions*, Yorick Wilks, ed., (Amsterdam, The Netherlands: John Benjamins, 2010), pp. 63-74; http://www.cs.bath.ac.uk/~jjb/ftp/Bryson-Slaves-Book09.html.

_______________, "Patiency Is Not a Virtue: AI and the Design of Ethical Systems," https://www.cs.bath.ac.uk/~jjb/ftp/Bryson-Patiency-AAAISS16.pdf.

So what we are creating are not—should not be—conscious, humanoid agents but an entirely new sort of entities, rather like oracles, with no conscience, no fear of death, no distracting loves and hates, no personality (but all sorts of foibles and quirks that would no doubt be identified as the "personality" of the system): boxes of truths (if we're lucky) almost certainly contaminated with a scattering of falsehoods. It will be hard enough learning to live with them without distracting ourselves with fantasies about the Singularity in which these AIs will enslave us, literally. The *human* use of human beings will soon be changed—once again—forever, but we can take the tiller and steer between some of the hazards if we take responsibility for our trajectory.

The roboticist **Rodney Brooks,** *featured in Errol Morris's 1997 documentary* Fast, Cheap and Out of Control *along with a lion-tamer, a topiarist, and an expert on the naked mole rat, was described by one reviewer as "smiling with a wild gleam in his eye." But that's pretty much true of most visionaries.*

A few years later in his career, Brooks, as befits one of the world's leading roboticists, suggested that "we overanthropomorphize humans, who are after all mere machines." He went on to present a warm-hearted vision of a coming AI world in which "the distinction between us and robots is going to disappear." He also admitted to something of a divided worldview. "Like a religious scientist, I maintain two sets of inconsistent beliefs and act on each of them in different circumstances," he wrote. "It is this transcendence between belief systems that I think will be what enables mankind to ultimately accept robots as emotional machines, and thereafter start to empathize with them and attribute free will, respect, and ultimately rights to them."

That was in 2002. In these pages, he takes a somewhat more jaundiced, albeit narrower, view; he is alarmed by the extent to which we have come to rely on pervasive systems that are not just exploitative but also vulnerable, as a result of the too-rapid development of software engineering—an advance that seems to have outstripped the imposition of reliably effective safeguards.

THE INHUMAN MESS OUR MACHINES HAVE GOTTEN US INTO

Rodney Brooks

***Rodney Brooks** is a computer scientist; Panasonic Professor of Robotics, emeritus, MIT; former director, MIT Computer Science Lab; and founder, chairman, and CTO of Rethink Robotics. He is the author of* Flesh and Machines.

Mathematicians and scientists are often limited in how they see the big picture, beyond their particular field, by the tools and metaphors they use in their work. Norbert Wiener is no exception, and I might guess that neither am I.

When he wrote *The Human Use of Human Beings,* Wiener was straddling the end of the era of understanding machines and animals simply as physical processes and the beginning of our current era of understanding machines and animals as computational processes. I suspect there will be future eras whose tools will look as distinct from the tools of the two eras Wiener straddled as those tools did from each other.

Wiener was a giant of the earlier era and built on the tools developed since the time of Newton and Leibniz to describe and analyze continuous processes in the physical world. In 1948 he published *Cybernetics,* a word he coined to describe the science of communication and control in both machines and animals. Today we would refer to the ideas in this book as control theory, an indispensable discipline for the design and analysis of physical machines, while mostly neglecting Wiener's claims about the science of communication. Wiener's innovations were largely driven by his work during the Second World War on mechanisms to aim and fire anti-aircraft guns. He brought mathematical rigor to the design of the sorts of technology whose design processes had been largely heuristic in nature: from the Roman waterworks through Watt's steam engine to the early development of automobiles.

One can imagine a different contingent version of our intellectual and technological history had Alan Turing and John von Neumann, both of whom made major contributions to the foundations of computing, not appeared on the scene. Turing contributed a fundamental model of computation—now known as a Turing Machine—in his paper "On Computable Numbers with an Application to the Entscheidungsproblem," written and revised in 1936 and published in 1937. In these machines, a linear tape of symbols from a finite alphabet encodes the input for a computational problem and also provides the working space for the computation. A different machine was required for each separate computational problem; later work by others would show that in one particular machine, now known as a Universal Turing Machine, an arbitrary set of computing instructions could be encoded on that same tape.

In the 1940s, von Neumann developed an abstract self-reproducing machine called a cellular automaton. In this case it occupied a finite subset of an infinite two-dimensional array of squares each containing a single symbol from a finite alphabet of twenty-nine distinct symbols—the rest of the infinite array starts out blank. The single symbols in each square change in lockstep, based on a complex but finite rule about the current symbol in that square and its immediate neighbors. Under the complex rule that von Neumann developed, most of the symbols in most of the squares stay the same and a few change at each step. So when one looks at the non-blank squares, it appears that there is a constant structure with some activity going on inside it. When von Neumann's abstract machine reproduced, it made a copy of itself in another region of the plane. Within the "machine" was a horizontal line of squares which acted as a finite linear tape, using a subset of the finite alphabet. It was the symbols in those squares that encoded the machine of which they were a part. During the machine's reproduction, the "tape" could move either left or right and was both interpreted (transcribed) as the instructions (translation) for the new "machine" being built and then copied (replicated)—with the new copy being placed inside the new machine for further reproduction. Francis Crick and James Watson later showed, in 1953, how such a tape could be instantiated in biology by a long DNA molecule with its finite alphabet of four nucleobases: guanine, cytosine, adenine, and thymine (G, C, A, and T).[13] As in von Neumann's machine, in biological reproduction the linear sequence of symbols in DNA is interpreted—through transcription into RNA molecules, which then are translated into proteins, the structures that make up a new cell—and the DNA is replicated and encased in the new cell.

A second foundational piece of work was in a 1945 "First Draft" report on the design for a digital computer, wherein von Neumann advocated for a memory that could contain both instructions and data.[14] This is now known as a von Neumann architecture computer—as distinct from a Harvard architecture computer, where there are two separate memories, one for instructions and one for data. The vast majority of computer chips built in the era of Moore's Law are based on the von Neumann architecture, including those powering our data centers, our laptops, and our smartphones. Von Neumann's digital-computer architecture is conceptually the same generalization—from early digital computers constructed with electromagnetic relays at both Harvard University and Bletchley Park—that occurs in going from a special-purpose Turing Machine to a Universal Turing Machine. Furthermore, his self-replicating automata share a fundamental similarity with both the construction of a Turing Machine and the mechanism of DNA-based reproducing biological cells. There is to this day scholarly debate over whether von Neumann saw the cross connections between these three pieces of work, Turing's and his two. Turing's revision of his paper was done while he and von Neumann were both at Princeton; indeed, after getting his PhD, Turing almost stayed on as von Neumann's postdoc.

Without Turing and von Neumann, the cybernetics of Wiener might have remained a dominant mode of thought and driver of technology for much longer than its brief moment of supremacy. In this imaginary version of history, we might well live today in an actual steam-punk world and not just get to observe its fantastical instantiations at Maker Faires!

My point is that Wiener thought about the world—physical, biological, and (in *Human Use*) sociological—in a particular way. He analyzed the world as continuous variables, as he explains in chapter 1 along with a nod to thermodynamics through an overlay of Gibbs statistics. He also shoehorns in a weak and unconvincing model of information as message-passing between and among both physical and biological entities. To me, and from today's vantage point seventy years on, his tools seem woefully inadequate for describing the mechanisms underlying biological systems, and so he missed out on how similar mechanisms might eventually be embodied in technological computational systems—as now they have been. Today's dominant technologies were developed in the world of Turing and von Neumann, rather than the world of Wiener.

---

[13] "A Structure for Deoxyribose Nucleic Acid," *Nature* 171, 737–738 (1953).

[14] https://en.wikipedia.org//wiki/First_Draft_of_a_Report_on_the_EDVAC#Controversy. Von Neumann is listed as the only author, whereas others contributed to the concepts he laid out; thus credit for the architecture has gone to him alone.

In the first industrial revolution, energy from a steam engine or a water wheel was used by human workers to replace their own energy. Instead of being a source of energy for physical work, people became modulators of how a large source of energy was used. But because steam engines and water wheels had to be large to be an efficient use of capital, and because in the 18th century the only technology for spatial distribution of energy was mechanical and worked only at very short range, many workers needed to be crowded around the source of energy. Wiener correctly argues that the ability to transmit energy as electricity caused a second industrial revolution. Now the source of energy could be distant from where it was used, and from the beginning of the 20th century, manufacturing could be much more dispersed as electrical-distribution grids were built.

Wiener then argues that a further new technology, that of the nascent computational machines of his time, will provide yet another revolution. The machines he talks about seem to be both analog and (perhaps) digital in nature; and he points out, in *The Human Use of Human Beings*, that since they will be able to make decisions, both blue-collar and white-collar workers may be reduced to being mere cogs in a much bigger machine. He fears that humans might use and abuse one another through organizational structures that this capability will encourage. We have certainly seen this play out in the last sixty years, and that disruption is far from over.

However, his physics-based view of computation protected him from realizing just how bad things might get. He saw machines' ability to communicate as providing a new and more inhuman way of exerting command and control. He missed that within a few decades computation systems would become more like biological systems, and it seems, from his descriptions in chapter 10 of his own work on modeling some aspects of biology, that he woefully underappreciated the many orders of magnitude of further complexity of biology over physics. We are in a much more complex situation today than he foresaw, and I am worried that it is much more pernicious than even his worst imagined fears.

In the 1960s, computation became firmly based on the foundations set out by Turing and von Neumann, and it was digital computation, based on the idea of finite alphabets which they both used. An arbitrarily long sequence, or string, formed by characters from a finite alphabet, can be encoded as a unique integer. As with Turing Machines themselves, the formalism for computation became that of computing an integer-valued function of a single integer-valued input.

Turing and von Neumann both died in the 1950s and at that time this is how they saw computation. Neither foresaw the exponential increase in computing capability that Moore's Law would bring—nor how pervasive computing machinery would become. Nor did they foresee two developments in our modeling of computation, each of which poses a great threat to human society.

The first is rooted in the abstractions they adopted. In the fifty-year, Moore's Law–fueled race to produce software that could exploit the doubling of computer capability every two years, the typical care and certification of engineering disciplines was thrown by the wayside. Software engineering was fast and prone to failures. This rapid development of software without standards of correctness has opened up many routes to exploit von Neumann architecture's storage of data and instructions in the same memory. One of the most common routes, known as "buffer overrun," involves an input number (or long string of characters) that is bigger than the programmer expected and overflows into where the instructions are stored. By carefully designing an input number that is too big by far, someone using a piece of software can infect it with instructions not intended by the programmer, and thus change what it does. This is the basis for creating a computer virus—so named for its similarity to a biological virus. The latter injects extra DNA into a cell, and that cell's transcription and translation mechanism blindly interprets it, making proteins that may be harmful to the host cell. Furthermore, the replication mechanism for the cell takes care of multiplying the virus. Thus, a small foreign entity can take control of a much bigger entity and bend its behavior in unexpected ways.

These and other forms of digital attacks have taken the security of our everyday lives from us. We rely on computers for almost everything now. We rely on computers for our infrastructure of electricity, gas, roads, cars, trains, and airplanes; these are all vulnerable. We rely on computers for our banking, our payment of bills, our retirement accounts, our mortgages, our purchasing of goods and services—these, too, are all vulnerable. We rely on computers for our entertainment, our communications both business and personal, our physical security at home, our information about the world, and our voting systems—all vulnerable. None of this will get fixed anytime soon. In the meantime, many aspects of our society are open to vicious attacks, whether by freelancing criminals or nation-state adversaries.

The second development is that computation has gone beyond simply computing functions. Instead, programs remain online continuously, and so they can gather data about a sequence of queries. Under the Wiener/Turing/von Neumann scheme, we might think of the communication pattern for a Web browser to be:

> *User: Give me Web page A.*
> *Browser: Here is Web page A.*
> …
> *User: Give me Web page B.*
> *Browser: Here is Web page B.*

Now instead it can look like this:

> *User: Give me Web page A.*
> *Browser: Here is Web page A. [And I will secretly remember that you asked for Web page A.]*
> …
> *User: Give me Web page B.*
> *Browser: Here is Web page B. [I see a correlation between its contents and that of the earlier requested Web page A, so I will update my model of you, the user, and transmit it to the company that produced me.]*

When the machine no longer simply computes a function but instead maintains a state, it can start to make inferences about the human by the sequence of requests presented to it. And when different programs correlate across different request streams—say, correlating Web-page searches with social-media posts, or the payment for services on another platform, or the dwell time on a particular advertisement, or where the user has walked or driven with their GPS-enabled smartphone—the total systems of many programs communicating with one another and with databases leads to a whole new loss of privacy. The great exploitative leap made by so many West Coast companies has been to monetize those inferences without the knowing permission of the person generating the interactions with the computing machine platforms.

Wiener, Turing, and von Neumann could not foresee the complexity of those platforms, wherein the legal mumbo-jumbo of the terms-of-use contracts the humans willingly enter into, without an inkling of what they entail, leads them to give up rights they would never concede in a one-on-one interaction with another human being. The computation platforms have become a shield behind which some companies hide in order to inhumanly exploit others. In certain other countries, the governments carry out these manipulations, and there the goal is not profits but the suppression of dissent.

Humankind has gotten itself into a fine pickle: We are being exploited by companies that paradoxically deliver services we crave, and at the same time our lives depend on many software-enabled systems that are open to attack. Getting ourselves out of this mess will be a long-term project. It will involve engineering, legislation, and most important, moral leadership. Moral leadership is the first and biggest challenge.

I first met **Frank Wilczek** *in the 1980s, when he invited me to his home in Princeton to talk about anyons. "The address is 112 Mercer Street," he wrote. "Look for the house with no driveway." So there I was, a few hours later, in Einstein's old living room, talking to a future recipient of the Nobel Prize in physics. If Frank was as impressed as I was by the surroundings, you'd never guess it. His only comment concerned the difficulty of finding a parking place in front of a "house with no driveway."*

Unlike most theoretical physicists, Frank has long had a keen interest in AI, as witnessed in these three "Observations":

1. "Francis Crick called it 'the Astonishing Hypothesis': that consciousness, also known as Mind, is an emergent property of matter," which, if true, indicates that "all intelligence is machine intelligence. What distinguishes natural from artificial intelligence is not what it is, but only how it is made."

2. "Artificial intelligence is not the product of an alien invasion. It is an artifact of a particular human culture and reflects the values of that culture."

3. "David Hume's striking statement 'Reason Is, and Ought only to Be, the Slave of the Passions' was written in 1738 [and] was, of course, meant to apply to human reason and human passions. . . . But Hume's logical/philosophical point remains valid for AI. Simply put: Incentives, not abstract logic, drive behavior."

He notes that "the big story of the 20th and the 21st century is that [as] computing develops, we learn how to calculate the consequences of the [fundamental] laws better and better. There's also a feedback cycle: When you understand matter better, you can design better computers, which will enable you to calculate better. It's kind of an ascending helix."

Here he argues that human intelligence, for now, holds the advantage—yet our future, unbounded by our solar system and doubtless also by our galaxy, will never be realized without the help of our AIs.

THE UNITY OF INTELLIGENCE

Frank Wilczek

***Frank Wilczek** is Herman Feshbach Professor of Physics at MIT, recipient of the 2004 Nobel Prize in physics, and the author of* A Beautiful Question: Finding Nature's Deep Design.

I. A Simple Answer to Contentious Questions:

- Can an artificial intelligence be conscious?
- Can an artificial intelligence be creative?
- Can an artificial intelligence be evil?

Those questions are often posed today, both in popular media and in scientifically informed debates. But the discussions never seem to converge. Here I'll begin by answering them as follows:

Based on physiological psychology, neurobiology, and physics, it would be very surprising if the answers were not Yes, Yes, and Yes. The reason is simple, yet profound: Evidence from those fields makes it overwhelmingly likely that there is no sharp divide between natural and artificial intelligence.

In his 1994 book of that title, the renowned biologist Francis Crick proposed an "astonishing hypothesis": that mind emerges from matter. He famously claimed that mind, in all its aspects, is "no more than the behavior of a vast assembly of nerve cells and their associated molecules."

The "astonishing hypothesis" is in fact the foundation of modern neuroscience. People try to understand how minds work by understanding how brains function; and they try to understand how brains function by studying how information is encoded in electrical and chemical signals, transformed by physical processes, and used to control behavior. In that scientific endeavor, they make no allowance for extraphysical behavior. So far, in thousands of exquisite experiments, that strategy has never failed. It has never proved necessary to allow for the influence of consciousness or creativity unmoored from brain activity to explain any observed fact of psychophysics or neurobiology. No one has ever stumbled upon a power of mind which is separate from conventional physical events in biological organisms. While there are many things we do not understand about brains, and about minds, the "astonishing hypothesis" has held intact.

If we broaden our view beyond neurobiology to consider the whole range of scientific experimentation, the case becomes still more compelling. In modern physics, the foci of interest are often extremely delicate phenomena. To investigate them, experimenters must take many precautions against contamination by "noise." They often find it necessary to construct elaborate shielding against stray electric and magnetic fields; to compensate for tiny vibrations due to micro-earthquakes or passing cars; to work at extremely low temperatures and in high vacuum, and so forth. But there's a notable exception: They have never found it necessary to make allowances for what people nearby (or, for that matter, far away) are thinking. No "thought waves," separate from known physical processes yet capable of influencing physical events, seem to exist.

That conclusion, taken at face value, erases the distinction between natural and artificial intelligence. It implies that if we were to duplicate, or accurately simulate, the physical processes occurring in a brain—as, in principle, we can—and wire up its input and output to sense organs and muscles, then we would reproduce, in a physical artifact, the observed manifestations of natural intelligence. Nothing observable would be missing. As an observer, I'd have no less (and no more) reason to ascribe consciousness, creativity, or evil to that artifact than I do to ascribe those properties to its natural counterparts, like other human beings.

Thus, by combining Crick's "astonishing hypothesis" in neurobiology with powerful evidence from physics, we deduce that natural intelligence is a special case of artificial intelligence. That conclusion deserves a name, and I will call it "the astonishing corollary."

With that, we have the answer to our three questions. Since consciousness, creativity, and evil are obvious features of natural human intelligence, they are possible features of artificial intelligence.

A hundred years ago, or even fifty, to believe the hypothesis that mind emerges from matter, and to infer our corollary that natural intelligence is a special case of artificial intelligence, would have been leaps of faith. In view of the many surrounding gaps—chasms, really—in contemporary understanding of biology and physics, they were genuinely doubtful propositions. But epochal developments in those areas have changed the picture:

In biology: A century ago, not only thought but also metabolism, heredity, and perception were deeply mysterious aspects of life that defied physical explanation. Today, of course, we have extremely rich and detailed accounts of metabolism, heredity, and many aspects of perception, from the bottom up, starting at the molecular level.

In physics: After a century of quantum physics and its application to materials, physicists have discovered, over and over, how rich and strange the behavior of matter can be. Superconductors, lasers, and many other wonders demonstrate that large assemblies of molecular units, each simple in itself, can exhibit qualitatively new, "emergent" behavior, while remaining fully obedient to the laws of physics. Chemistry, including biochemistry, is a cornucopia of emergent phenomena, all now quite firmly grounded in physics. The pioneering physicist Philip Anderson, in an essay titled "More Is Different," offers a classic discussion of emergence. He begins by acknowledging that "the reductionist hypothesis [i.e., the completeness of physical explanations based on known interactions of simple parts] may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted without question." But he goes on to emphasize that "[t]he behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles."[15] Each new level of size and complexity supports new forms of organization, whose patterns encode information in new ways and whose behavior is best described using new concepts.

Electronic computers are a magnificent example of emergence. Here, all the cards are on the table. Engineers routinely design, from the bottom up, based on known (and quite sophisticated) physical principles, machines that process information in extremely impressive ways. Your iPhone can beat you at chess, quickly collect and deliver information about anything, and take great pictures, too. Because the process whereby computers, smartphones, and other intelligent objects are designed and manufactured is completely transparent, there can be no doubt that their wonderful capabilities emerge from regular physical processses, which we can trace down to the level of electrons, photons, quarks, and gluons. Evidently, brute matter can get pretty smart.

---

[15] *Science,* 4 August 1972, Vol. 177, No. 4047, pp. 393-96.

Let me summarize the argument. From two strongly supported hypotheses, we've drawn a straightforward conclusion:

- Human mind emerges from matter.
- Matter is what physics says it is.
- *Therefore,* the human mind emerges from physical processes we understand and can reproduce artificially.
- *Therefore,* natural intelligence is a special case of artificial intelligence.

Of course, our "astonishing corollary" could fail; the first two lines of this argument are hypotheses. But their failure would have to bring in a foundation-shattering discovery—a significant new phenomenon, with large-scale physical consequences, which takes place in unremarkable, well-studied physical circumstances (i.e., the materials, temperatures, and pressures inside human brains) yet which has somehow managed for many decades to elude determined investigators armed with sophisticated instruments. Such a discovery would be. . . astonishing.

II. The Future of Intelligence

It is part of human nature to improve on human bodies and minds. Historically, clothing, eyeglasses, and watches are examples of increasingly sophisticated augmentations that enhance our toughness, perception, and awareness. They are major improvements to the natural human endowment, whose familiarity should not blind us to their depth. Today smartphones and the Internet are bringing the human drive toward augmentation into realms more central to our identity as intelligent beings. They are giving us, in effect, quick access to a vast collective awareness and a vast collective memory.

At the same time, autonomous artificial intelligences have become world champions in a wide variety of "cerebral" games, such as chess and Go, and have taken over many sophisticated pattern-recognition tasks, such as reconstructing what happened during complex reactions at the Large Hadron Collider from a blizzard of emerging particle tracks, to find new particles; or gathering clues from fuzzy X-ray, fMRI, and other types of images, to diagnose medical problems.

Where is this drive toward self-enhancement and innovation taking us? While the precise sequence of events and the timescale over which they'll play out is impossible to predict (or, at least, beyond me), some basic considerations suggest that eventually the most powerful embodiments of mind will be quite different things from human brains as we know them today.

Consider six factors whereby information-processing technology exceeds human capabilities—vastly, qualitatively, or both:

- *Speed:* The orchestrated motion of electrons, which is the heart of modern artificial information-processing, can be much faster than the processes of diffusion and chemical change by which brains operate. Typical modern computer clock rates approach 10 gigahertz, corresponding to 10 billion operations per second. No single measure of speed applies to the bewildering variety of brain processes, but one fundamental limitation is latency of action potentials, which limits their spacing to a few 10s per second. It is probably no accident that the "frame rate," at which we can distinguish that movies are actually a sequence of stills, is about 40 per second. Thus, electronic processing is close to a billion times faster.

- *Size:* The linear dimension of a typical neuron is about 10 microns. Molecular dimensions, which set a practical limit, are about 10,000 times smaller, and artificial processing units are approaching that scale. Smallness makes communication more efficient.
- *Stability:* Whereas human memory is essentially continuous (analog), artificial memory can incorporate discrete (digital) features. Whereas analog quantities can erode, digital quantities can be stored, refreshed, and maintained with complete accuracy.
- *Duty Cycle:* Human brains grow tired with effort. They need time off to take nourishment and to sleep. They carry the burden of aging. Most profoundly: They die.
- *Modularity* (open architecture): Because artificial information processors can support precisely defined digital interfaces, they can readily assimilate new modules. Thus, if we want a computer to "see" ultraviolet or infrared or "hear" ultrasound, we can feed the output from an appropriate sensor directly into its "nervous system." The architecture of brains is much more closed and opaque, and the human immune system actively resists implants.
- *Quantum readiness*: One case of modularity deserves special mention, because of its long-term potential. Recently physicists and information scientists have come to appreciate that the principles of quantum mechanics support new computing principles, which can empower qualitatively new forms of information processing and (plausibly) new levels of intelligence. But these possibilities rely on aspects of quantum behavior which are quite delicate and seem especially unsuitable for interfacing with the warm, wet, messy enviroment of human brains.

Evidently, as platforms for intelligence, human brains are far from optimal. Still, although versatile housekeeping robots or mechanical soldiers would find ready, lucrative markets, at present there is no machine that approaches the kind of general-purpose human intelligence those applications would require. Despite their relative weakness on many fronts, human brains have some big advantages over their artificial competitors. Let me mention five:

- *Three-dimensionality:* Although, as noted, the linear dimensions of existing artificial processing units are vastly smaller than those of brains, the procedure by which they're made—centered on lithography (basically, etching)—is essentially two-dimensional. That is revealed visibly in the geometry of computer boards and chips. Of course, one can stack boards, but the spacing between layers is much larger, and communication much less efficient, than within layers. Brains make better use of all three dimensions.
- *Self-repair:* Human brains can recover from, or work around, many kinds of injuries or errors. Computers often must be repaired or rebooted externally.

- *Connectivity:* Human neurons typically support several hundred connections (synapses). Moreover, the complex pattern of these connections is very meaningful. (See our next point.) Computer units typically make only a handful of connections, in regular, fixed patterns.
- *Development* (self-assembly with interactive sculpting): The human brain grows its units by cell divisions and orchestrates them into coherent structures by movement and folding. It also proliferates an abundance of connections among the cells. An important part of its sculpting occurs through active processes during infancy and childhood, as the individual interacts with his or her environment. In this process, many connections are winnowed away, while others are strengthened, depending on their effectiveness in use. Thus, the fine structure of the brain is tuned through interaction with the external world—a rich source of information and feedback!
- *Integration* (sensors and actuators): The human brain comes equipped with a variety of sensory organs, notably including its outgrowth eyes, and with versatile actuators, including hands that build, legs that walk, and mouths that speak. Those sensors and actuators are seamlessy integrated into the brain's information-processing centers, having been honed over millions of years of natural selection. We interpret their raw signals and control their large-scale actions with minimal conscious attention. The flip side is that we don't know how we do it, and the implementation is opaque. It's proving surprisingly difficult to reach human standards on these "routine" input-output functions.

These advantages of human brains over currently engineered artifacts are profound. Human brains supply an inspiring existence proof, showing us several ways we can get more out of matter. When, if ever, will our engineering catch up?

I don't know for sure, but let me offer some informed opinions. The challenges of three-dimensionality and, to a lesser extent, self-repair don't look overwhelming. They present some tough engineering problems, but many incremental improvements are easy to imagine, and there are clear paths forward. And while the powers of human eyes, hands, and other sensory organs and actuators are wonderfully effective, their abilities are far from exhausting any physical limits. Optical systems can take pictures with higher resolution in space, time, and color, and in more regions of the electromagnetic spectrum; robots can move faster and be stronger; and so forth. In these domains, the components necessary for superhuman performance, along many axes, are already available. The bottleneck is getting information into and out of them, rapidly, in the language of the information-processing units.

And this brings us to the remaining, and I think most profound, advantages of brains over artificial devices, which stem from their connectivity and interactive development. Those two advantages are synergistic, since it is interactive development that sculpts the massively wired but sprawling structure of the infant brain, enabled by exponential growth of neurons and synapses, to get tuned into the extraordinary instrument it becomes. Computer scientists are beginning to discover the power of the brain's architecture: Neural nets, whose basic design, as their name suggests, was directly inspired by the brain's, have scored some spectacular successes in game playing and pattern recognition, as noted. But present-day engineering has nothing comparable—in the (currently) esoteric domain of self-reproducing machines—to the power and versatility of neurons and their synapses. This could become a new, great frontier of research. Here too, biology might point the way, as we come to understand biological development well enough to imitate its essence.

Altogether, the advantages of artificial over natural intelligence appear permanent, while the advantages of natural over artificial intelligence, though substantial at present, appear transient. I'd guess that it will be many decades before engineering catches up, but—barring catastrophic wars, climate change, or plagues, so that technological progress stays vigorous—few centuries.

If that's right, we can look forward to several generations during which humans, empowered and augmented by smart devices, coexist with increasingly capable autonomous AIs. There will be a complex, rapidly changing ecology of intelligence, and rapid evolution in consequence. Given the intrinsic advantages that engineered devices will eventually offer, the vanguard of that evolution will be cyborgs and superminds, rather than lightly adorned *Homo sapiens.*

Another important impetus will come from the exploration of hostile environments, both on Earth (e.g., the deep ocean) and, especially, in space. The human body is poorly adapted to conditions outside a narrow band of temperatures, pressures, and atmospheric composition. It needs a wide variety of specific, complex nutrients, and plenty of water. Also, it is not radiation-hardened. As the manned space program has amply demonstrated, it is difficult and expensive to maintain humans outside their terrestrial comfort zone. Cyborgs or autonomous AIs could be much more effective in these explorations. Quantum AIs, with their sensitivity to noise, might even be happier in the cold and dark of deep space.

In a moving passage from his 1935 novel *Odd John,* science fiction's singular genius Olaf Stapledon has his hero, a superhuman (mutant) intelligence, describe *Homo sapiens* as "the Archaeopteryx of the spirit." He says this, fondly, to his friend and biographer, who is a normal human. Archaeopteryx was a noble creature, and a bridge to greater ones.

I was introduced to **Max Tegmark** *some years ago by his MIT colleague Alan Guth, the father of the inflationary universe. A distinguished theoretical physicist and cosmologist himself, Max's principal concern nowadays is the looming existential risk posed by the creation of an AGI (artificial general intelligence—that is, one that matches human intelligence). Four years ago, Max co-founded, with Jaan Tallinn and others, the Future of Life Institute (FLI), which bills itself as "an outreach organization working to ensure that tomorrow's most powerful technologies are beneficial for humanity." While on a book tour in London, he was in the midst of planning for FLI, and he admits being driven to tears in a tube station after a trip to the London Science Museum, with its exhibitions spanning the gamut of humanity's technological achievements. Was all that impressive progress in vain?*

FLI's scientific advisory board includes Elon Musk, Frank Wilczek, George Church, Stuart Russell, and the Oxford philosopher Nick Bostrom, who dreamed up an oft-quoted Gedankenexperiment *that results in a world full of paper clips and nothing else, produced by an (apparently) well-meaning AGI who was just following orders. The Institute sponsors conferences (Puerto Rico 2015, Asilomar 2017) on AI safety issues and in 2018 instituted a grants competition focusing on research in aid of maximizing the societal benefits of AGI.*

While Max is sometimes listed—by the non-cognoscenti—on the side of the scaremongers, he believes, like Frank Wilczek, in a future that will immensely benefit from AGI if, in the attempt to create it, we can keep the human species from being sidelined.

LET'S ASPIRE TO MORE THAN MAKING OURSELVES OBSOLETE

Max Tegmark

***Max Tegmark** is an MIT physicist and AI researcher; president of the Future of Life Institute; scientific director of the Foundational Questions Institute; and the author of* Our Mathematical Universe *and* Life 3.0: Being Human in the Age of Artificial Intelligence.

Although there's great controversy about how and when AI will impact humanity, the situation is clearer from a cosmic perspective: The technology-developing life that has evolved on Earth is rushing to make itself obsolete without devoting much serious thought to the consequences. This strikes me as embarrassingly lame, given that we can create amazing opportunities for humanity to flourish like never before, if we dare to steer a more ambitious course.

13.8 billion years after its birth, our Universe has become aware of itself. On a small blue planet, tiny conscious parts of our Universe have discovered that what they once thought was the sum total of existence was a minute part of something far grander: a solar system in a galaxy in a universe with over 100 billion other galaxies, arranged into an elaborate pattern of groups, clusters, and superclusters.

Consciousness is the cosmic awakening; it transformed our Universe from a mindless zombie with no self-awareness into a living ecosystem harboring self-reflection, beauty, hope, meaning, and purpose. Had that awakening never taken place, our Universe would have been pointless—a gigantic waste of space. Should our Universe go back to sleep permanently due to some cosmic calamity or self-inflicted mishap, it will become meaningless again.

On the other hand, things could get even better. We don't yet know whether we humans are the only stargazers in the cosmos, or even the first, but we've already learned enough about our Universe to know that it has the potential to wake up much more fully than it has thus far. AI pioneers such as Norbert Wiener have taught us that a further awakening of our Universe's ability to process and experience information need not require eons of additional evolution but perhaps mere decades of human scientific ingenuity.

We may be like that first glimmer of self-awareness you experienced when you emerged from sleep this morning, a premonition of the much greater consciousness that would arrive once you opened your eyes and fully awoke. Perhaps artificial superintelligence will enable life to spread throughout the cosmos and flourish for billions or trillions of years, and perhaps this will be because of decisions we make here, on our planet, in our lifetime.

Or humanity may soon go extinct, through some self-inflicted calamity caused by the power of our technology growing faster than the wisdom with which we manage it.

The evolving debate about AI's societal impact

Many thinkers dismiss the idea of superintelligence as science fiction, because they view intelligence as something mysterious that can exist only in biological organisms—especially humans—and as fundamentally limited to what today's humans can do. But from my perspective as a physicist, intelligence is simply a certain kind of information processing performed by elementary particles moving around, and there's no law of physics that says one can't build machines more intelligent in every way than we are, and able to seed cosmic life. This suggests that we've seen just the tip of the intelligence iceberg; there's an amazing potential to unlock the full intelligence latent in nature and use it to help humanity flourish—or flounder.

Others, including some of the authors in this volume, dismiss the building of an AGI (Artificial General Intelligence—an entity able to accomplish any cognitive task at least as well as humans) not because they consider it physically impossible but because they deem it too difficult for humans to pull off in less than a century. Among professional AI researchers, both types of dismissal have become minority views because of recent breakthroughs. There is a strong expectation that AGI will be achieved within a century, and the median forecast is only decades away. A recent survey of AI researchers by Vincent Muller and Nick Bostrom concludes:

> [T]he results reveal a view among experts that AI systems will probably (over 50%) reach overall human ability by 2040-50, and very likely (with 90% probability) by 2075. From reaching human ability, it will move on to superintelligence in 2 years (10%) to 30 years (75%) thereafter.[16]

In the cosmic perspective of gigayears, it makes little difference whether AGI arrives in thirty or three hundred years, so let's focus on the implications rather than the timing.

First, we humans discovered how to replicate some natural processes with machines, making our own heat, light, and mechanical horsepower. Gradually we realized that our bodies were also machines, and the discovery of nerve cells blurred the boundary between body and mind. Finally, we started building machines that could outperform not only our muscles but our minds as well. We've now been eclipsed by machines in the performance of many narrow cognitive tasks, ranging from memorization and arithmetic to game play, and we are in the process of being overtaken in many more, from driving to investing to medical diagnosing. If the AI community succeeds in its original goal of building AGI, then we will have, by definition, been eclipsed at all cognitive tasks.

This begs many obvious questions. For example, will whoever or whatever controls the AGI control Earth? Should we aim to control superintelligent machines? If not, can we ensure that they understand, adopt, and retain human values? As Norbert Wiener put it in *The Human Use of Human Beings:*

> Woe to us if we let [the machine] decide our conduct, unless we have previously examined the laws of its action, and know fully that its conduct will be carried out on principles acceptable to us! On the other hand, the machine . . . , which can learn and can make decisions on the basis of its learning, will in no way be obliged to make such decisions as we should have made, or will be acceptable to us.

---

[16] Vincent C. Müller & Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," in *Fundamental Issues of Artificial Intelligence*, Vincent C. Muller, ed. (Springer International Publishing Switzerland, 2016), pp. 555-72. https://nickbostrom.com/papers/survey.pdf.

And who are the "us"? Who should deem "such decisions . . . acceptable"? Even if future powers decide to help humans survive and flourish, how will we find meaning and purpose in our lives if we aren't needed for anything?

The debate about the societal impact of AI has changed dramatically in the last few years. In 2014, what little public talk there was of AI risk tended to be dismissed as Luddite scaremongering, for one of two logically incompatible reasons:

(1) AGI was overhyped and wouldn't happen for at least another century.

(2) AGI would probably happen sooner but was virtually guaranteed to be beneficial.

Today, talk of AI's societal impact is everywhere, and work on AI safety and AI ethics has moved into companies, universities, and academic conferences. The controversial position on AI safety research is no longer to advocate for it but to dismiss it. Whereas the open letter that emerged from the 2015 Puerto Rico AI conference (and helped mainstream AI safety) spoke only in vague terms about the importance of keeping AI beneficial, the 2017 Asilomar AI Principles (see below) had real teeth: They explicitly mention recursive self-improvement, superintelligence, and existential risk, and were signed by AI industry leaders and over a thousand AI researchers from around the world.

Nonetheless, most discussion is limited to the near-term impact of narrow AI and the broader community pays only limited attention to the dramatic transformations that AGI may soon bring to life on Earth. Why?

Why we're rushing to make ourselves obsolete, and why we avoid talking about it

First of all, there's simple economics. Whenever we figure out how to make another type of human work obsolete by building machines that do it better and cheaper, most of society gains: Those who build and use the machines make profits, and consumers get more affordable products. This will be as true of future investor AGIs and scientist AGIs as it was of weaving machines, excavators, and industrial robots. In the past, displaced workers usually found new jobs, but this basic economic incentive will remain even if that is no longer the case. The existence of affordable AGI means, by definition, that *all* jobs can be done more cheaply by machines, so anyone claiming that "people will always find new well-paying jobs" is in effect claiming that AI researchers will fail to build AGI.

Second, *Homo sapiens* is by nature curious, which will motivate the scientific quest for understanding intelligence and developing AGI even without economic incentives. Although curiosity is one of the most celebrated human attributes, it can cause problems when it fosters technology we haven't yet learned how to manage wisely. Sheer scientific curiosity without profit motive contributed to the discovery of nuclear weapons and tools for engineering pandemics, so it's not unthinkable that the old adage "Curiosity killed the cat" will turn out to apply to the human species as well.

Third, we're mortal. This explains the near unanimous support for developing new technologies that help us live longer, healthier lives, which strongly motivates current AI research. AGI can clearly aid medical research even more. Some thinkers even aspire to near immortality via cyborgization or uploading.

We're thus on the slippery slope toward AGI, with strong incentives to keep sliding downward, even though the consequence will by definition be our economic obsolescence. We will no longer be needed for anything, because all jobs can be done more efficiently by machines. The successful creation of AGI would be the biggest event in human history, so why is there so little serious discussion of what it might lead to?

Here again, the answer involves multiple reasons.

First, as Upton Sinclair famously quipped, "It is difficult to get a man to understand something, when his salary depends on his not understanding it."[17] For example, spokesmen for tech companies or university research groups often claim there are no risks attached to their activities even if they privately think otherwise. Sinclair's observation may help explain not only reactions to risks from smoking and climate change but also why some treat technology as a new religion whose central articles of faith are that more technology is always better and whose heretics are clueless scaremongering Luddites.

Second, humans have a long track record of wishful thinking, flawed extrapolation of the past, and underestimation of emerging technologies. Darwinian evolution endowed us with powerful fear of concrete threats, not of abstract threats from future technologies that are hard to visualize or even imagine. Consider trying to warn people in 1930 of a future nuclear arms race, when you couldn't show them a single nuclear explosion video and nobody even knew how to build such weapons. Even top scientists can underestimate uncertainty, making forecasts that are either too optimistic—Where are those fusion reactors and flying cars?—or too pessimistic. Ernest Rutherford, arguably the greatest nuclear physicist of his time, said in 1933—less than twenty-four hours before Leo Szilard conceived of the nuclear chain reaction—that nuclear energy was "moonshine." Essentially nobody at that time saw the nuclear arms race coming.

Third, psychologists have discovered that we tend to avoid thinking of disturbing threats when we believe there's nothing we can do about them anyway. In this case, however, there are many constructive things we can do, if we can get ourselves to start thinking about the issue.

What can we do?

I'm advocating a strategy change from *"Let's rush to build technology that makes us obsolete—what could possibly go wrong?"* to *"Let's envision an inspiring future and steer toward it."*

To motivate the effort required for steering, this strategy begins by envisioning an enticing destination. Although Hollywood's futures tend to be dystopian, the fact is that AGI can help life flourish as never before. Everything I love about civilization is the product of intelligence, so if we can amplify our own intelligence with AGI, we have the potential to solve today's and tomorrow's thorniest problems, including disease, climate change, and poverty. The more detailed we can make our shared positive visions for the future, the more motivated we will be to work together to realize them.

What should we do in terms of steering? The twenty-three Asilomar principles adopted in 2017 offer plenty of guidance, including these short-term goals:

(1) An arms race in lethal autonomous weapons should be avoided.

(2) The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

---

[17] Upton Sinclair, *I, Candidate for Governor: And How I Got Licked* (Berkeley CA: University of California Press, 1994), p. 109.

(3) Investments in AI should be accompanied by funding for research on ensuring its beneficial use. . . . How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked.[18]

The first two involve not getting stuck in suboptimal Nash equilibria. An out-of-control arms race in lethal autonomous weapons that drives the price of automated anonymous assassination toward zero will be very hard to stop once it gains momentum. The second goal would require reversing the current trend in some Western countries where sectors of the population are getting poorer in absolute terms, fueling anger, resentment, and polarization. Unless the third goal can be met, all the wonderful AI technology we create might harm us, either accidentally or deliberately.

AI safety research must be carried out with a strict deadline in mind: Before AGI arrives, we need to figure out how to make AI understand, adopt, and retain our goals. The more intelligent and powerful machines get, the more important it becomes to align their goals with ours. As long as we build relatively dumb machines, the question isn't whether human goals will prevail but merely how much trouble the machines can cause before we solve the goal-alignment problem. If a superintelligence is ever unleashed, however, it will be the other way around: Since intelligence is the ability to accomplish goals, a superintelligent AI is by definition much better at accomplishing its goals than we humans are at accomplishing ours, and will therefore prevail.

In other words, *the real risk with AGI isn't malice but competence*. A superintelligent AGI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours, we're in trouble. People don't think twice about flooding anthills to build hydroelectric dams, so let's not place humanity in the position of those ants. Most researchers argue that if we end up creating superintelligence, we should make sure it's what AI-safety pioneer Eliezer Yudkowsky has termed "friendly AI"—AI whose goals are in some deep sense beneficial.

The moral question of what these goals should be is just as urgent as the technical questions about goal alignment. For example, what sort of society are we hoping to create, where we find meaning and purpose in our lives even though we, strictly speaking, aren't needed? I'm often given the following glib response to this question: "*Let's build machines that are smarter than us and then let them figure out the answer!*" This mistakenly equates intelligence with morality. Intelligence isn't good or evil but morally neutral. It's simply an ability to accomplish complex goals, good or bad. We can't conclude that things would have been better if Hitler had been more intelligent. Indeed, postponing work on ethical issues until after goal-aligned AGI is built would be irresponsible and potentially disastrous. A perfectly obedient superintelligence whose goals automatically align with those of its human owner would be like Nazi SS-Obersturmbannführer Adolf Eichmann on steroids. Lacking moral compass or inhibitions of its own, it would, with ruthless efficiency, implement its owner's goals, whatever they might be.[19]

When I speak of the need to analyze technology risk, I'm sometimes accused of scaremongering. But here at MIT, where I work, we know that such risk analysis isn't scaremongering: It's *safety engineering*. Before the moon-landing mission, NASA systematically thought through everything that could possibly go wrong when putting astronauts on top of a 110-meter rocket full of highly flammable fuel and launching them to a place where nobody could help them—and there were lots of things that could go wrong. Was this scaremongering? No, this was the safety engineering that ensured the mission's success. Similarly, we should analyze what could go wrong with AI to ensure that it goes right.

---

[18] https://futureoflife.org/ai-principles/

[19] See, for example, Hannah Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil* (New York: Penguin Classics, 2006).

Outlook

In summary, if our technology outpaces the wisdom with which we manage it, it can lead to our extinction. It's already caused the extinction of from 20 to 50 percent of all species on Earth, by some estimates,[20] and it would be ironic if we're next in line. It would also be pathetic, given that the opportunities offered by AGI are literally astronomical, potentially enabling life to flourish for billions of years not only on Earth but also throughout much of our cosmos.

Instead of squandering this opportunity through unscientific risk denial and poor planning, let's be ambitious! *Homo sapiens* is inspiringly ambitious, as reflected in William Ernest Henley's famous lines from *Invictus*: "I am the master of my fate, / I am the captain of my soul." Rather than drifting like a rudderless ship toward our own obsolescence, let's take on and overcome the technical and societal challenges standing between us and a good high-tech future. What about the existential challenges related to morality, goals, and meaning? There's no meaning encoded in the laws of physics, so instead of passively waiting for our Universe to give meaning to us, let's acknowledge and celebrate that it's we conscious beings who give meaning to our Universe. Let's create our own meaning, based on something more profound than having jobs. AGI can enable us to finally become the masters of our own destiny. Let's make that destiny a truly inspiring one!

---

[20] See Elizabeth Kolbert, *The Sixth Extinction: An Unnatural History* (New York: Henry Holt, 2014).

Jaan Tallinn *grew up in Estonia, becoming one of its few computer game developers, when that nation was still a Soviet Socialist Republic. Here he compares the dissidents who brought down the Iron Curtain to the dissidents who are sounding the alarm about rapid advances in artificial intelligence. He locates the roots of the current AI dissidence, paradoxically, among such pioneers of the AI field as Wiener, Alan Turing, and I. J. Good.*

Jaan's preoccupation is with existential risk, AI being among the most extreme of many. In 2012, he co-founded the Centre for the Study of Existential Risk—an interdisciplinary research institute that works to mitigate risks "associated with emerging technologies and human activity"—at the University of Cambridge, along with philosopher Huw Price and Martin Rees, the Astronomer Royal.

He once described himself to me as "a convinced consequentialist"—convinced enough to have given away much of his entrepreneurial wealth to the Future of Life Institute (of which he is a co-founder), the Machine Intelligence Research Institute, and other such organizations working on risk reduction. Max Tegmark has written about him: "If you're an intelligent life-form reading this text millions of years from now and marveling at how life is flourishing, you may owe your existence to Jaan."

On a recent visit to London, Jaan and I participated on an AI panel for the Serpentine Gallery's Marathon at London's City Hall, under the aegis of Hans Ulrich Obrist (another contributor to this volume). This being the art world, there was a glamorous dinner party that night in a mansion filled with London's beautiful people—artists, fashion models, oligarchs, stars of stage and screen. After working the room in his unaffected manner ("Hi, I'm Jaan"), he suddenly said, "Time for hip-hop dancing," dropped to the floor on one hand, and began demonstrating his spectacular moves to the bemused A-listers. Then off he went into the dance-club subculture, which is apparently how he ends every evening when he's on the road. Who knew?

DISSIDENT MESSAGES
Jaan Tallinn

***Jaan Tallin**, a computer programmer, theoretical physicist, and investor, is a co-developer of Skype and Kazaa.*

In March 2009, I found myself in a bland franchise eatery next to a noisy California freeway. I was there to meet a young man whose blog I had been following. To make himself recognizable, he wore a button with a text on it: *Speak the truth even if your voice trembles*. His name was Eliezer Yudkowsky, and we spent the next four hours discussing the message he had for the world—a message that had brought me to that eatery and would end up dominating my subsequent work.

The First Message: the Soviet Occupation

In *The Human Use of Human Beings*, Norbert Wiener looked at the world through the lens of communication. He saw a universe that was marching to the tune of the second law of thermodynamics toward its inevitable heat death. In such a universe, the only (meta)stable entities are *messages*—patterns of information that propagate through time, like waves propagating across the surface of a lake. Even we humans can be considered messages, because the atoms in our bodies are too fleeting to attach our identities to. Instead, we are the "message" that our bodily functions maintain. As Wiener put it: "It is the pattern maintained by this homeostasis, which is the touchstone of our personal identity."

I'm more used to treating processes and computation as the fundamental building blocks of the world. That said, Wiener's lens brings out some interesting aspects of the world which might otherwise have remained in the background and which to a large degree shaped my life. These are two messages, both of which have their roots in the Second World War. They started out as quiet dissident messages—messages that people didn't pay much attention to, even if they silently and perhaps subconsciously concurred. The first message was: *The Soviet Union is composed of a series of illegitimate occupations. These occupations must end.*

As an Estonian, I grew up behind the Iron Curtain and had a front row seat when it fell. I heard this first message in the nostalgic reminiscences of my grandparents and in between the harsh noises jamming the Voice of America. It grew louder during the Gorbachev era, as the state became more lenient in its treatment of dissidents, and reached a crescendo in the Estonian Singing Revolution of the late 1980s.

In my teens, I witnessed the message spread out across widening circles of people, starting with the active dissidents, who had voiced it for half a century at great cost to themselves, proceeding to the artists and literati, and ending up among the Party members and politicians who had switched sides. This new elite comprised an eclectic mix of people: those original dissidents who had managed to survive the repression, public intellectuals, and (to the great annoyance of the surviving dissidents) even former Communists. The remaining dogmatists—even the prominent ones—were eventually marginalized, some of them retreating to Russia.

Interestingly, as the message propagated from one group to the next, it evolved. It started in pure and uncompromising form ("The occupation must end!") among the dissidents who considered the truth more important than their personal freedom. The mainstream groups, who had more to lose, initially qualified and diluted the message, taking positions like, "It would make sense in the long term to delegate control over local matters." (There were always exceptions: Some public intellectuals proclaimed the original dissident message verbatim.) Finally, the original message—being, simply, true—won out over its diluted versions. Estonia regained its independence in 1991, and the last Soviet troops left three years later.

The people who took the risk and spoke the truth in Estonia and elsewhere in the Eastern Bloc played a monumental role in the eventual outcome—an outcome that changed the lives of hundreds of millions of people, myself included. They spoke the truth, even as their voices trembled.

The Second Message: AI Risk

My exposure to the second revolutionary message was via Yudkowsky's blog—the blog that compelled me to reach out and arrange that meeting in California. The message was: *Continued progress in AI can precipitate a change of cosmic proportions—a runaway process that will likely kill everyone. We need to put in a lot of extra effort to avoid that outcome*.

After my meeting with Yudkowsky, the first thing I did was try to interest my Skype colleagues and close collaborators in his warning. I failed. The message was too crazy, too dissident. Its time had not yet come.

Only later did I learn that Yudkowsky wasn't the original dissident speaking this particular truth. In April 2000, there was a lengthy opinion piece in *Wired* titled, "Why the Future Doesn't Need Us," by Bill Joy, co-founder and chief scientist of Sun Microsystems. He warned:

> Accustomed to living with almost routine scientific breakthroughs, we have yet to come to terms with the fact that the most compelling 21st-century technologies—robotics, genetic engineering, and nanotechnology—pose a different threat than the technologies that have come before. Specifically, robots, engineered organisms, and nanobots share a dangerous amplifying factor: They can self-replicate. . . . [O]ne bot can become many, and quickly get out of control.

Apparently, Joy's broadside caused a lot of furor but little action.

More surprising to me, though, was that the AI-risk message arose almost simultaneously with the field of computer science. In a 1951 lecture, Alan Turing announced: "[I]t seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. . . . At some stage, therefore, we should have to expect the machines to take control. . . ."[21] A decade or so later, his Bletchley Park colleague I. J. Good wrote, "The first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."[22] Indeed, I counted half a dozen places in *The Human Use of Human Beings* where Wiener hinted at one or another aspect of the Control Problem. ("The machine like the djinnee, which can learn and can make decisions on the basis of its learning, will in no way be obliged to make such decisions as we should have made, or will be acceptable to us.") Apparently, the original dissidents promulgating the AI-risk message were the AI pioneers themselves!

---

[21] Posthumously reprinted in *Phil. Math.* (3) vol. 4, 256-60 (1966).

[22] Irving John Good, "Speculations concerning the first ultraintelligent machine," *Advances in Computers,* vol. 6 (Academic Press, 1965), pp. 31-88.

Evolution's Fatal Mistake

There have been many arguments, some sophisticated and some less so, for why the Control Problem is real and not some science-fiction fantasy. Allow me to offer one that illustrates the magnitude of the problem:

For the last hundred thousand years, the world (meaning the Earth, but the argument extends to the solar system and possibly even to the entire universe) has been in the human-brain regime. In this regime, the brains of *Homo sapiens* have been the most sophisticated future-shaping mechanisms (indeed, some have called them the most complicated objects in the universe). Initially, we didn't use them for much beyond survival and tribal politics in a band of foragers, but now their effects are surpassing those of natural evolution. The planet has gone from producing forests to producing cities.

As predicted by Turing, once we have superhuman AI ("the machine thinking method"), the human-brain regime will end. Look around you—you're witnessing the final decades of a hundred-thousand-year regime. This thought alone should give people some pause before they dismiss AI as just another tool. One of the world's leading AI researchers recently confessed to me that he would be greatly relieved to learn that human-level AI was impossible for us to create.

Of course, it might still take us a long time to develop human-level AI. But we have reason to suspect that this is not the case. After all, it didn't take long, in relative terms, for evolution—the blind and clumsy optimization process—to create human-level intelligence once it had animals to work with. Or multicellular life, for that matter: Getting cells to stick together seems to have been much harder for evolution to accomplish than creating humans once there were multicellular organisms. Not to mention that our level of intelligence was limited by such grotesque factors as the width of the birth canal. Imagine an AI developer being stopped in his tracks because he couldn't manage to adjust the font size on his computer!

There's an interesting symmetry here: In fashioning humans, evolution created a system that is, at least in many important dimensions, a more powerful planner and optimizer than evolution itself is. We are the first species to understand that we're the product of evolution. Moreover, we've created many artifacts (radios, firearms, spaceships) that evolution would have little hope of creating. Our future, therefore, will be determined by our own decisions and no longer by biological evolution. In that sense, evolution has fallen victim to its own Control Problem.

We can only hope that we're smarter than evolution in that sense. We *are* smarter, of course, but will that be enough? We're about to find out.

The Present Situation

So here we are, more than half a century after the original warnings by Turing, Wiener, and Good, and a decade after people like me started paying attention to the AI-risk message. I'm glad to see that we've made a lot of progress in confronting this issue, but we're definitely not there yet. AI risk, although no longer a taboo topic, is not yet fully appreciated among AI researchers. AI risk is not yet common knowledge either. In relation to the timeline of the first dissident message, I'd say we're around the year 1988, when raising the Soviet-occupation topic was no longer a career-ending move but you still had to somewhat hedge your position. I hear similar hedging now—statements like, "I'm not concerned about superintelligent AI, but there are some real ethical issues in increased automation," or "It's good that some people are researching AI risk, but it's not a short-term concern," or even the very reasonable sounding, "These are small-probability scenarios, but their potentially high impact justifies the attention."

As far as message propagation goes, though, we are getting close to the tipping point. A recent survey of AI researchers who published at the two major international AI conferences in 2015 found that 40 percent now think that risks from highly advanced AI are either "an important problem" or "among the most important problems in the field."[23]

Of course, just as there were dogmatic Communists who never changed their position, it's all but guaranteed that some people will never admit that AI is potentially dangerous. Many of the deniers of the first kind came from the Soviet nomenklatura; similarly, the AI-risk deniers often have financial or other pragmatic motives. One of the leading motives is corporate profits. AI is profitable, and even in instances where it isn't, it's at least a trendy, forward-looking enterprise with which to associate your company. So a lot of the dismissive positions are products of corporate PR and legal machinery. In some very real sense, big corporations are nonhuman machines that pursue their own interests—interests that might not align with those of any particular human working for them. As Wiener observed in *The Human Use of Human Beings*: "When human atoms are knit into an organization in which they are used, not in their full right as responsible human beings, but as cogs and levers and rods, it matters little that their raw material is flesh and blood."

Another strong incentive to turn a blind eye to the AI risk is the (very human) curiosity that knows no bounds. "When you see something that is technically sweet, you go ahead and do it and you argue about what to do about it only after you have had your technical success. That is the way it was with the atomic bomb," said J. Robert Oppenheimer. His words were echoed recently by Geoffrey Hinton, arguably the inventor of deep learning, in the context of AI risk: "I could give you the usual arguments, but the truth is that the prospect of discovery is too sweet."

Undeniably, we have both entrepreneurial attitude and scientific curiosity to thank for almost all the nice things we take for granted in the modern era. It's important to realize, though, that progress does not *owe* us a good future. In Wiener's words, "It is possible to believe in progress as a fact without believing in progress as an ethical principle."

Ultimately, we don't have the luxury of waiting before all the corporate heads and AI researchers are willing to concede the AI risk. Imagine yourself sitting in a plane about to take off. Suddenly there's an announcement that 40 percent of the experts believe there's a bomb onboard. At that point, the course of action is already clear, and sitting there waiting for the remaining 60 percent to come around isn't part of it.

---

[23] Katja Grace, et al., "When Will AI Exceed Human Performance? Evidence from AI Experts," https://arxiv.org/pdf/1705.08807.pdf.

Calibrating the AI-Risk Message

While uncannily prescient, the AI-risk message from the original dissidents has a giant flaw—as does the version dominating current public discourse: Both considerably understate the magnitude of the problem as well as AI's potential upside. The message, in other words, does not adequately convey the stakes of the game.

Wiener primarily warned of the *social* risks—risks stemming from careless integration of machine-generated decisions with governance processes and misuse (by humans) of such automated decision making. Likewise, the current "serious" debate about AI risks focuses mostly on things like technological unemployment or biases in machine learning. While such discussions can be valuable and address pressing short-term problems, they are also stunningly parochial. I'm reminded of Yudkowsky's quip in a blog post: "[A]sking about the effect of *machine superintelligence* on the conventional human labor market is like asking how US–Chinese trade patterns would be affected by the Moon crashing into the Earth. There would indeed be effects, but you'd be missing the point."

In my view, the central point of the AI risk is that *superintelligent AI is an environmental risk*. Allow me to explain.

In his "Parable of the Sentient Puddle," Douglas Adams describes a puddle that wakes up in the morning and finds himself in a hole that fits him "staggeringly well." From that observation, the puddle concludes that the world must have been made for him. Therefore, writes Adams, "the moment he disappears catches him rather by surprise." To assume that AI risks are limited to adverse social developments is to make a similar mistake. The harsh reality is that the universe was not made for us; instead, we are fine-tuned by evolution to a very narrow range of environmental parameters. For instance, we need the atmosphere at ground level to be roughly at room temperature, at about 100 kPa pressure, and have a sufficient concentration of oxygen. Any disturbance, even temporary, of this precarious equilibrium and we die in a matter of minutes.

Silicon-based intelligence does not share such concerns about the environment. That's why it's much cheaper to explore space using machine probes rather than "cans of meat." Moreover, Earth's current environment is almost certainly suboptimal for what a superintelligent AI will greatly care about: *efficient computation*. Hence we might find our planet suddenly going from anthropogenic global warming to machinogenic global cooling. One big challenge that AI safety research needs to deal with is how to constrain a potentially superintelligent AI—an AI with a much larger footprint than our own—from rendering our environment uninhabitable for biological life-forms.

Interestingly, given that the most potent sources both of AI research and AI-risk dismissals are under big corporate umbrellas, if you squint hard enough the "AI as an environmental risk" message looks like the chronic concern about corporations skirting their environmental responsibilities.

Conversely, the worry about AI's social effects also misses most of the upside. It's hard to overemphasize how tiny and parochial the future of our planet is, compared with the full potential of humanity. On astronomical timescales, our planet will be gone soon (unless we tame the sun, also a distinct possibility) and almost all the resources—atoms and free energy—to sustain civilization in the long run are in deep space.

Eric Drexler, the inventor of nanotechnology, has recently been popularizing the concept of “Pareto-topia”: the idea that AI, if done right, can bring about a future in which *everyone's* lives are hugely improved, a future where there are no losers. A key realization here is that what chiefly prevents humanity from achieving its full potential might be our instinctive sense that we’re in a zero-sum game—a game in which players are supposed to eke out small wins at the expense of others. Such an instinct is seriously misguided and destructive in a “game” where everything is at stake and the payoff is literally astronomical. There are many more star systems in our galaxy alone than there are people on Earth.

Hope

As of this writing, I’m cautiously optimistic that the AI-risk message can save humanity from extinction, just as the Soviet-occupation message ended up liberating hundreds of millions of people. As of 2015, it had reached and converted 40 percent of AI researchers. It wouldn’t surprise me if a new survey now would show that the majority of AI researchers believe AI safety to be an important issue.

I’m delighted to see the first technical AI-safety papers coming out of DeepMind, OpenAI, and Google Brain and the collaborative problem-solving spirit flourishing between the AI-safety research teams in these otherwise very competitive organizations.

The world’s political and business elite are also slowly waking up: AI safety has been covered in reports and presentations by the Institute of Electrical and Electronics Engineers (IEEE), the World Economic Forum, and the Organization for Economic Cooperation and Development (OECD). Even the recent (July 2017) Chinese AI manifesto contained dedicated sections on “AI safety supervision” and “Develop[ing] laws, regulations, and ethical norms” and establishing “an AI security and evaluation system” to, among other things, “[e]nhance the awareness of risk.” I very much hope that a new generation of leaders who understand the AI Control Problem and AI as the ultimate environmental risk can rise above the usual tribal, zero-sum games and steer humanity past these dangerous waters we are in—thereby opening our way to the stars that have been waiting for us for billions of years.

Here’s to our next hundred thousand years! And don’t hesitate to speak the truth, even if your voice trembles.

Throughout his career, whether studying language, advocating a realistic biology of mind, or examining the human condition through the lens of humanistic Enlightenment ideas, psychologist **Steven Pinker** *has embraced and championed a naturalistic understanding of the universe and the computational theory of mind. He is perhaps the first internationally recognized public intellectual whose recognition is based on the advocacy of empirically based thinking about language, mind, and human nature.*

"Just as Darwin made it possible for a thoughtful observer of the natural world to do without creationism," he says, "Turing and others made it possible for a thoughtful observer of the cognitive world to do without spiritualism."

In the debate about AI risk, he argues against prophecies of doom and gloom, noting that they spring from the worst of our psychological biases—exemplified particularly by media reports: "Disaster scenarios are cheap to play out in the probability-free zone of our imaginations, and they can always find a worried, technophobic, or morbidly fascinated audience." Hence, over the centuries: Pandora, Faust, the Sorcerer's Apprentice, Frankenstein, the population bomb, resource depletion, HAL, suitcase nukes, the Y2K bug, and engulfment by nanotechnological grey goo. "A characteristic of AI dystopias," he points out, "is that they project a parochial alpha-male psychology onto the concept of intelligence. . . . History does turn up the occasional megalomaniacal despot or psychopathic serial killer, but these are products of a history of natural selection shaping testosterone-sensitive circuits in a certain species of primate, not an inevitable feature of intelligent systems."

In the present essay, he applauds Wiener's belief in the strength of ideas vis-à-vis the encroachment of technology. As Wiener so aptly put it, "The machine's danger to society is not from the machine itself but from what man makes of it."

TECH PROPHECY AND THE UNDERAPPRECIATED CAUSAL POWER OF IDEAS

Steven Pinker

***Steven Pinker,** a Johnstone Family Professor in the Department of Psychology at Harvard University, is an experimental psychologist who conducts research in visual cognition, psycholinguistics, and social relations. He is the author of eleven books, including* The Blank Slate, The Better Angels of Our Nature, *and, most recently,* Enlightenment Now: The Case for Reason, Science, Humanism, and Progress.

Artificial intelligence is an existence proof of one of the great ideas in human history: that the abstract realm of knowledge, reason, and purpose does not consist of an élan vital or immaterial soul or miraculous powers of neural tissue. Rather, it can be linked to the physical realm of animals and machines via the concepts of information, computation, and control. Knowledge can be explained as patterns in matter or energy that stand in systematic relations with states of the world, with mathematical and logical truths, and with one another. Reasoning can be explained as transformations of that knowledge by physical operations that are designed to preserve those relations. Purpose can be explained as the control of operations to effect changes in the world, guided by discrepancies between its current state and a goal state. Naturally evolved brains are just the most familiar systems that achieve intelligence through information, computation, and control. Humanly designed systems that achieve intelligence vindicate the notion that information processing is sufficient to explain it—the notion that the late Jerry Fodor dubbed the computational theory of mind.

The touchstone for this volume, Norbert Wiener's *The Human Use of Human Beings,* celebrated this intellectual accomplishment, of which Wiener himself was a foundational contributor. A potted history of the mid-20th-century revolution that gave the world the computational theory of mind might credit Claude Shannon and Warren Weaver for explaining knowledge and communication in terms of information. It might credit Alan Turing and John von Neumann for explaining intelligence and reasoning in terms of computation. And it ought to give Wiener credit for explaining the hitherto mysterious world of purposes, goals, and teleology in terms of the technical concepts of feedback, control, and cybernetics (in its original sense of "governing" the operation of a goal-directed system). "It is my thesis," he announced, "that the physical functioning of the living individual and the operation of some of the newer communication machines are precisely parallel in their analogous attempts to control entropy through feedback"—the staving off of life-sapping entropy being the ultimate goal of human beings.

Wiener applied the ideas of cybernetics to a third system: society. The laws, norms, customs, media, forums, and institutions of a complex community could be considered channels of information propagation and feedback that allow a society to ward off disorder and pursue certain goals. This is a thread that runs through the book and which Wiener himself may have seen as its principal contribution. In his explanation of feedback, he wrote, "This complex of behavior is ignored by the average man, and in particular does not play the role that it should in our habitual analysis of society; for just as individual physical responses may be seen from this point of view, so may the organic responses of society itself."

Indeed, Wiener gave scientific teeth to the idea that in the workings of history, politics, and society, *ideas matter*. Beliefs, ideologies, norms, laws, and customs, by regulating the behavior of the humans who share them, can shape a society and power the course of historical events as surely as the phenomena of physics affect the structure and evolution of the solar system. To say that ideas—and not just weather, resources, geography, or weaponry—can shape history is not woolly mysticism. It is a statement of the causal powers of information instantiated in human brains and exchanged in networks of communication and feedback. Deterministic theories of history, whether they identify the causal engine as technological, climatological, or geographic, are belied by the causal power of ideas. The effects of these ideas can include unpredictable lurches and oscillations that arise from positive feedback or from miscalibrated negative feedback.

An analysis of society in terms of its propagation of ideas also gave Wiener a guideline for social criticism. A healthy society—one that gives its members the means to pursue life in defiance of entropy—allows information sensed and contributed by its members to feed back and affect how the society is governed. A dysfunctional society invokes dogma and authority to impose control from the top down. Wiener thus described himself as "a participant in a liberal outlook," and devoted most of the moral and rhetorical energy in the book (both the 1950 and 1954 editions) to denouncing communism, fascism, McCarthyism, militarism, and authoritarian religion (particularly Catholicism and Islam) and to warning that political and scientific institutions were becoming too hierarchical and insular.

Wiener's book is also, here and there, an early exemplar of an increasingly popular genre, tech prophecy. Prophecy not in the sense of mere prognostications but in the Old Testament sense of dark warnings of catastrophic payback for the decadence of one's contemporaries. Wiener warned against the accelerating nuclear arms race, against technological change that was imposed without regard to human welfare ("[W]e must know as scientists what man's nature is and what his built-in purposes are"), and against what today is called the value-alignment problem: that "the machine like the djinnee, which can learn and can make decisions on the basis of its learning, will in no way be obliged to make such decisions as we should have made, or will be acceptable to us." In the darker, 1950 edition, he warned of a "threatening new Fascism dependent on the *machine à gouverner*."

Wiener's tech prophecy harks back to the Romantic movement's rebellion against the "dark Satanic mills" of the Industrial Revolution, and perhaps even earlier, to the archetypes of Prometheus, Pandora, and Faust. And today it has gone into high gear. Jeremiahs, many of them (like Wiener) from the worlds of science and technology, have sounded alarms about nanotechnology, genetic engineering, Big Data, and particularly artificial intelligence. Several contributors to this volume characterize Wiener's book as a prescient example of tech prophecy and amplify his dire worries.

Yet the two moral themes of *The Human Use of Human Beings*—the liberal defense of an open society and the dystopian dread of runaway technology—are in tension. A society with channels of feedback that maximize human flourishing will have mechanisms in place, and can adapt them to changing circumstances, in a way that can domesticate technology to human purposes. There's nothing idealistic or mystical about this; as Wiener emphasized, ideas, norms, and institutions are themselves a form of technology, consisting of patterns of information distributed across brains. The possibility that machines threaten a new fascism must be weighed against the vigor of the liberal ideas, institutions, and norms that Wiener championed throughout the book. The flaw in today's dystopian prophecies is that they disregard the existence of these norms and institutions, or drastically underestimate their causal potency. The result is a technological determinism whose dark predictions are repeatedly refuted by the course of events. The numbers "1984" and "2001" are good reminders.

I will consider two examples. Tech prophets often warn of a "surveillance state" in which a government empowered by technology will monitor and interpret all private communications, allowing it to detect dissent and subversion as it arises and make resistance to state power futile. Orwell's telescreens are the prototype, and in 1976 Joseph Weizenbaum, one of the gloomiest tech prophets of all time, warned my class of graduate students not to pursue automatic speech recognition because government surveillance was its only conceivable application.

Though I am on record as an outspoken civil libertarian, deeply concerned with contemporary threats to free speech, I lose no sleep over technological advances in the Internet, video, or artificial intelligence. The reason is that almost all the variation across time and space in freedom of thought is driven by differences in norms and institutions and almost none of it by differences in technology. Though one can imagine hypothetical combinations of the most malevolent totalitarians with the most advanced technology, in the real world it's the norms and laws we should be vigilant about, not the tech.

Consider variation across time. If, as Orwell hinted, advancing technology was a prime enabler of political repression, then Western societies should have gotten more and more restrictive of speech over the centuries, with a dramatic worsening in the second half of the 20th century continuing into the 21st. That's not how history unfolded. It was the centuries when communication was implemented by quills and inkwells that had autos-da-fé and the jailing or guillotining of Enlightenment thinkers. During World War I, when the state of the art was the wireless, Bertrand Russell was jailed for his pacifist opinions. In the 1950s, when computers were room-size accounting machines, hundreds of liberal writers and scholars were professionally punished. Yet in the technologically accelerating, hyperconnected 21st century, 18 percent of social science professors are Marxists[24]; the President of the United States is nightly ridiculed by television comedians as a racist, pervert, and moron; and technology's biggest threat to political discourse comes from amplifying too many dubious voices rather than suppressing enlightened ones.

Now consider variations across place. Western countries at the technological frontier consistently get the highest scores in indexes of democracy and human rights, while many backward strongman states are at the bottom, routinely jailing or killing government critics. The lack of a correlation between technology and repression is unsurprising when you analyze the channels of information flow in any human society. For dissidents to be influential, they have to get their message out to a wide network via whatever channels of communication are available—pamphleteering, soap-box oration, subversive soirées in cafés and pubs, word of mouth. These channels enmesh influential dissidents in a broad social network which makes them easy to identify and track down.

---

[24] Neil Gross & Solon Simmons, "The Social and Political Views of American College and University Professors," in N. Gross & S. Simmons, eds., *Professors and Their Politics* (Baltimore: Johns Hopkins University Press, 2014).

All the more so when dictators rediscover the time-honored technique of weaponizing the people against each other by punishing those who don't denounce or punish others.

In contrast, technologically advanced societies have long had the means to install Internet-connected, government-monitored surveillance cameras in every bar and bedroom. Yet that has not happened, because democratic governments (even the current American administration, with its flagrantly antidemocratic impulses) lack the will and the means to enforce such surveillance on an obstreperous people accustomed to saying what they want. Occasionally, warnings of nuclear, biological, or cyberterrorism goad government security agencies into measures such as hoovering up mobile phone metadata, but these ineffectual measures, more theater than oppression, have had no significant effect on either security or freedom. Ironically, tech prophecy plays a role in encouraging these measures. By sowing panic about supposed existential threats such as suitcase nuclear bombs and bioweapons assembled in teenagers' bedrooms, they put pressure on governments to prove they're doing something, anything, to protect the American people.

It's not that political freedom takes care of itself. It's that the biggest threats lie in the networks of ideas, norms, and institutions that allow information to feed back (or not) on collective decisions and understanding. As opposed to the chimerical technological threats, one real threat today is oppressive political correctness, which has choked the range of publicly expressible hypotheses, terrified many intelligent people against entering the intellectual arena, and triggered a reactionary backlash. Another real threat is the combination of prosecutorial discretion with an expansive lawbook filled with vague statutes. The result is that every American unwittingly commits "three felonies a day" (as the title of a book by civil libertarian Harvey Silverglate puts it) and is in jeopardy of imprisonment whenever it suits the government's needs. It's this prosecutorial weaponry that makes Big Brother all-powerful, not telescreens. The activism and polemicizing directed against government surveillance programs would be better directed at its overweening legal powers.

The other focus of much tech prophecy today is artificial intelligence, whether in the original sci-fi dystopia of computers running amok and enslaving us in an unstoppable quest for domination, or the newer version in which they subjugate us by accident, single-mindedly seeking some goal we give them regardless of its side effects on human welfare (the value-alignment problem adumbrated by Wiener). Here again both threats strike me as chimerical, growing from a narrow technological determinism that neglects the networks of information and control in an intelligent system like a computer or brain and in a society as a whole.

The subjugation fear is based on a muzzy conception of intelligence that owes more to the Great Chain of Being and a Nietzschean will to power than to a Wienerian analysis of intelligence and purpose in terms of information, computation, and control. In these horror scenarios, intelligence is portrayed as an all-powerful, wish-granting potion that agents possess in different amounts. Humans have more of it than animals, and an artificially intelligent computer or robot will have more of it than humans. Since we humans have used our moderate endowment to domesticate or exterminate less well-endowed animals (and since technologically advanced societies have enslaved or annihilated technologically primitive ones), it follows that a supersmart AI would do the same to us. Since an AI will think millions of times faster than we do, and use its superintelligence to recursively improve its superintelligence, from the instant it is turned on we will be powerless to stop it.

But these scenarios are based on a confusion of intelligence with motivation—of beliefs with desires, inferences with goals, the computation elucidated by Turing and the control elucidated by Wiener. Even if we did invent superhumanly intelligent robots, why would they *want* to enslave their masters or take over the world? Intelligence is the ability to deploy novel means to attain a goal. But the goals are extraneous to the intelligence: Being smart is not the same as wanting something. It just so happens that the intelligence in *Homo sapiens* is a product of Darwinian natural selection, an inherently competitive process. In the brains of that species, reasoning comes bundled with goals such as dominating rivals and amassing resources. But it's a mistake to confuse a circuit in the limbic brain of a certain species of primate with the very nature of intelligence. There is no law of complex systems that says that intelligent agents must turn into ruthless megalomaniacs.

A second misconception is to think of intelligence as a boundless continuum of potency, a miraculous elixir with the power to solve any problem, attain any goal. The fallacy leads to nonsensical questions like when an AI will "exceed human-level intelligence," and to the image of an "artificial general intelligence" (AGI) with God-like omniscience and omnipotence. Intelligence is a contraption of gadgets: software modules that acquire, or are programmed with, knowledge of how to pursue various goals in various domains. People are equipped to find food, win friends and influence people, charm prospective mates, bring up children, move around in the world, and pursue other human obsessions and pastimes. Computers may be programmed to take on some of these problems (like recognizing faces), not to bother with others (like charming mates), and to take on still other problems that humans can't solve (like simulating the climate or sorting millions of accounting records). The problems are different, and the kinds of knowledge needed to solve them are different.

But instead of acknowledging the centrality of knowledge to intelligence, the dystopian scenarios confuse an artificial general intelligence of the future with Laplace's demon, the mythical being that knows the location and momentum of every particle in the universe and feeds them into equations for physical laws to calculate the state of everything at any time in the future. For many reasons, Laplace's demon will never be implemented in silicon. A real-life intelligent system has to acquire information about the messy world of objects and people by engaging with it one domain at a time, the cycle being governed by the pace at which events unfold in the physical world. That's one of the reasons that understanding does not obey Moore's Law: Knowledge is acquired by formulating explanations and testing them against reality, not by running an algorithm faster and faster. Devouring the information on the Internet will not confer omniscience either: Big Data is still finite data, and the universe of knowledge is infinite.

A third reason to be skeptical of a sudden AI takeover is that it takes too seriously the inflationary phase in the AI hype cycle in which we are living today. Despite the progress in machine learning, particularly multilayered artificial neural networks, current AI systems are nowhere near achieving general intelligence (if that concept is even coherent). Instead, they are restricted to problems that consist of mapping well-defined inputs to well-defined outputs in domains where gargantuan training sets are available, in which the metric for success is immediate and precise, in which the environment doesn't change, and in which no stepwise, hierarchical, or abstract reasoning is necessary. Many of the successes come not from a better understanding of the workings of intelligence but from the brute-force power of faster chips and Bigger Data, which allow the programs to be trained on millions of examples and generalize to similar new ones. Each system is an idiot savant, with little ability to leap to problems it was not set up to solve, and a brittle mastery of those it was. And to state the obvious, none of these programs has made a move toward taking over the lab or enslaving its programmers.

Even if an artificial intelligence system tried to exercise a will to power, without the cooperation of humans it would remain an impotent brain in a vat. A superintelligent system, in its drive for self-improvement, would somehow have to build the faster processors that it would run on, the infrastructure that feeds it, and the robotic effectors that connect it to the world—all impossible unless its human victims worked to give it control of vast portions of the engineered world. Of course, one can always imagine a Doomsday Computer that is malevolent, universally empowered, always on, and tamperproof. The way to deal with this threat is straightforward: Don't build one.

What about the newer AI threat, the value-alignment problem, foreshadowed in Wiener's allusions to stories of the Monkey's Paw, the genie, and King Midas, in which a wisher rues the unforeseen side effects of his wish? The fear is that we might give an AI system a goal and then helplessly stand by as it relentlessly and literal-mindedly implemented its interpretation of that goal, the rest of our interests be damned. If we gave an AI the goal of maintaining the water level behind a dam, it might flood a town, not caring about the people who drowned. If we gave it the goal of making paper clips, it might turn all the matter in the reachable universe into paper clips, including our possessions and bodies. If we asked it to maximize human happiness, it might implant us all with intravenous dopamine drips, or rewire our brains so we were happiest sitting in jars, or, if it had been trained on the concept of happiness with pictures of smiling faces, tile the galaxy with trillions of nanoscopic pictures of smiley-faces.

Fortunately, these scenarios are self-refuting. They depend on the premises that (1) humans are so gifted that they can design an omniscient and omnipotent AI, yet so idiotic that they would give it control of the universe without testing how it works; and (2) the AI would be so brilliant that it could figure out how to transmute elements and rewire brains, yet so imbecilic that it would wreak havoc based on elementary blunders of misunderstanding. The ability to choose an action that best satisfies conflicting goals is not an add-on to intelligence that engineers might forget to install and test; it *is* intelligence. So is the ability to interpret the intentions of a language user in context.

When we put aside fantasies like digital megalomania, instant omniscience, and perfect knowledge and control of every particle in the universe, artificial intelligence is like any other technology. It is developed incrementally, designed to satisfy multiple conditions, tested before it is implemented, and constantly tweaked for efficacy and safety.

The last criterion is particularly significant. The culture of safety in advanced societies is an example of the humanizing norms and feedback channels that Wiener invoked as a potent causal force and advocated as a bulwark against the authoritarian or exploitative implementation of technology. Whereas at the turn of the 20th century Western societies tolerated shocking rates of mutilation and death in industrial, domestic, and transportation accidents, over the course of the century the value of human life increased. As a result, governments and engineers used feedback from accident statistics to implement countless regulations, devices, and design changes that made technology progressively safer. The fact that some regulations (such as using a cell phone near a gas pump) are ludicrously risk-averse underscores the point that we have become a society obsessed with safety, with fantastic benefits as a result: Rates of industrial, domestic, and transportation fatalities have fallen by more than 95 (and often 99) percent since their highs in the first half of the 20th century.[25] Yet tech prophets of malevolent or oblivious artificial intelligence write as if this momentous transformation never happened and one morning engineers will hand total control of the physical world to untested machines, heedless of the human consequences.

Norbert Wiener explained ideas, norms, and institutions in terms of computational and cybernetic processes that were scientifically intelligible and causally potent. He explained human beauty and value as "a local and temporary fight against the Niagara of increasing entropy" and expressed the hope that an open society, guided by feedback on human well-being, would enhance that value. Fortunately his belief in the causal power of ideas counteracted his worries about the looming threat of technology. As he put it, "the machine's danger to society is not from the machine itself but from what man makes of it." It is only by remembering the causal power of ideas that we can accurately assess the threats and opportunities presented by artificial intelligence today.

---

[25] Steven Pinker, "Safety," *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress* (New York: Penguin, 2018).

The most significant developments in the sciences today (i.e., those that affect the lives of everybody on the planet) are about, informed by, or implemented through advances in software and computation. Central to the future of these developments is physicist **David Deutsch,** *the founder of the field of quantum computation, whose 1985 paper on universal quantum computers was the first full treatment of the subject; the Deutsch-Jozsa algorithm was the first quantum algorithm to demonstrate the enormous potential power of quantum computation.*

When he initially proposed it, quantum computation seemed practically impossible. But the explosion in the construction of simple quantum computers and quantum communication systems never would have taken place without his work. He has made many other important contributions in areas such as quantum cryptography and the many-worlds interpretation of quantum theory. In a philosophic paper (with Artur Ekert), he appealed to the existence of a distinctive quantum theory of computation to argue that our knowledge of mathematics is derived from, and subordinate to, our knowledge of physics (even though mathematical truth is independent of physics).

Because he has spent a good part of his working life changing people's worldviews, his recognition among his peers as an intellectual goes well beyond his scientific achievement. He argues (following Karl Popper) that scientific theories are "bold conjectures," not derived from evidence but only tested by it. His two main lines of research at the moment—qubit-field theory and constructor theory—may well yield important extensions of the computational idea.

In the following essay, he more or less aligns himself with those who see human-level artificial intelligence as promising us a better world rather than the Apocalypse. In fact, he pleads for AGI to be, in effect, given its head, free to conjecture—a proposition that several other contributors to this book would consider dangerous.

BEYOND REWARD AND PUNISHMENT
David Deutsch

***David Deutsch** is a quantum physicist and a member of the Centre for Quantum Computation at the Clarendon Laboratory, Oxford University. He is the author of* The Fabric of Reality *and* The Beginning of Infinity.

First Murderer:
We are men, my liege.
Macbeth:
Ay, in the catalogue ye go for men,
As hounds and greyhounds, mongrels, spaniels, curs,
Shoughs, water-rugs, and demi-wolves are clept
All by the name of dogs.
William Shakespeare – *Macbeth*

For most of our species' history, our ancestors were barely people. This was not due to any inadequacy in their brains. On the contrary, even before the emergence of our anatomically modern human sub-species, they were making things like clothes and campfires, using knowledge that was not in their genes. It was created in their brains by thinking, and preserved by individuals in each generation imitating their elders. Moreover, this must have been knowledge in the sense of *understanding*, because it is impossible to imitate novel complex behaviors like those without understanding what the component behaviors are for.[26]

Such knowledgeable imitation depends on successfully guessing explanations, whether verbal or not, of what the other person is trying to achieve and how each of his actions contributes to that—for instance, when he cuts a groove in some wood, gathers dry kindling to put in it, and so on.

The complex cultural knowledge that this form of imitation permitted must have been extraordinarily useful. It drove rapid evolution of anatomical changes, such as increased memory capacity and more gracile (less robust) skeletons, appropriate to an ever more technology-dependent lifestyle. No nonhuman ape today has this ability to imitate novel complex behaviors. Nor does any present-day artificial intelligence. But our pre-*sapiens* ancestors did.

Any ability based on guessing must include means of correcting one's guesses, since most guesses will be wrong at first. (There are always many more ways of being wrong than right.) Bayesian updating is inadequate, because it cannot generate novel guesses about the purpose of an action, only fine-tune—or, at best, choose among—existing ones. Creativity is needed. As the philosopher Karl Popper explained, creative criticism, interleaved with creative conjecture, is how humans learn one another's behaviors, including language, and extract meaning from one another's utterances.[27]

---

[26] "Aping" (imitating certain behaviors without understanding) uses inborn hacks such as the mirror-neuron system. But behaviors imitated that way are drastically limited in complexity. See Richard Byrne, "Imitation as Behaviour Parsing," *Phil. Trans. R. Soc.,* B 358:1431, 529-36 (2003).

[27] Karl Popper, *Conjectures and Refutations* (1963).

Those are also the processes by which all new knowledge is created: They are how we innovate, make progress, and create abstract understanding for its own sake. This is human-level intelligence: thinking. It is also, or should be, the property we seek in artificial general intelligence (AGI). Here I'll reserve the term "thinking" for processes that can create understanding (explanatory knowledge). Popper's argument implies that all thinking entities—human or not, biological or artificial—must create such knowledge in fundamentally the same way. Hence understanding any of those entities requires traditionally human concepts such as culture, creativity, disobedience, and morality—which justifies using the uniform term *people* to refer to all of them.

Misconceptions about human thinking and human origins are causing corresponding misconceptions about AGI and how it might be created. For example, it is generally assumed that the evolutionary pressure that produced modern humans was provided by the benefits of having an ever greater ability to innovate. But if that were so, there would have been rapid progress as soon as thinkers existed, just as we hope will happen when we create artificial ones. If thinking had been commonly used for anything other than imitating, it would also have been used for innovation, even if only by accident, and innovation would have created opportunities for further innovation, and so on exponentially. But instead, there were hundreds of thousands of years of near stasis. Progress happened only on timescales much longer than people's lifetimes, so in a typical generation no one benefited from any progress. Therefore, the benefits of the ability to innovate can have exerted little or no evolutionary pressure during the biological evolution of the human brain. That evolution was driven by the benefits of *preserving* cultural knowledge.

Benefits to the genes, that is. Culture, in that era, was a very mixed blessing to individual people. Their cultural knowledge was indeed good enough to enable them to outclass all other large organisms (they rapidly became the top predator, etc.), even though it was still extremely crude and full of dangerous errors. But culture consists of transmissible information—memes—and meme evolution, like gene evolution, tends to favor high-fidelity transmission. And high-fidelity meme transmission necessarily entails the suppression of attempted progress. So it would be a mistake to imagine an idyllic society of hunter-gatherers, learning at the feet of their elders to recite the tribal lore by heart, being content despite their lives of suffering and grueling labor and despite expecting to die young and in agony of some nightmarish disease or parasite. Because, even if they could conceive of nothing better than such a life, those torments were the least of their troubles. For suppressing innovation in human minds (without killing them) is a trick that can be achieved only by human action, and it is an ugly business.

This has to be seen in perspective. In the civilization of the West today, we are shocked by the depravity of, for instance, parents who torture and murder their children for not faithfully enacting cultural norms. And even more by societies and subcultures where that is commonplace and considered honorable. And by dictatorships and totalitarian states that persecute and murder entire harmless populations for behaving differently. We are ashamed of our own recent past, in which it was honorable to beat children bloody for mere disobedience. And before that, to own human beings as slaves. And before that, to burn people to death for being infidels, to the applause and amusement of the public. Steven Pinker's book *The Better Angels of our Nature* contains accounts of horrendous evils that were normal in historical civilizations. Yet even they did not extinguish innovation as efficiently as it was extinguished among our forebears in prehistory for thousands of centuries.[28]

That is why I say that prehistoric people, at least, were barely people. Both before and after becoming perfectly human both physiologically and in their mental potential, they were monstrously inhuman in the actual content of their thoughts. I'm not referring to their crimes or even their cruelty as such: Those are all too human. Nor could mere cruelty have reduced progress that effectively. Things like "the thumbscrew and the stake / For the glory of the Lord" [29] were for reining in the few deviants who had somehow escaped mental standardization, which would normally have taken effect long before they were in danger of inventing heresies. From the earliest days of thinking onward, children must have been cornucopias of creative ideas and paragons of critical thought—otherwise, as I said, they could not have learned language or other complex culture. Yet, as Jacob Bronowski stressed in *The Ascent of Man*:

> For most of history, civilisations have crudely ignored that enormous potential. . . . [C]hildren have been asked simply to conform to the image of the adult. . . . The girls are little mothers in the making. The boys are little herdsmen. They even carry themselves like their parents.

But of course, they weren't just "asked" to ignore their enormous potential and conform faithfully to the image fixed by tradition: They were somehow trained to be psychologically unable to deviate from it. By now, it is hard for us even to conceive of the kind of relentless, finely tuned oppression required to reliably extinguish, in everyone, the aspiration to progress and replace it with dread and revulsion at any novel behavior. In such a culture, there can have been no morality other than conformity and obedience, no other identity than one's status in a hierarchy, no mechanisms of cooperation other than punishment and reward. So everyone had the same aspiration in life: to avoid the punishments and get the rewards. In a typical generation, no one invented anything, because no one aspired to anything new, because everyone had already despaired of improvement being possible. Not only was there no technological innovation or theoretical discovery, there were no new worldviews, styles of art, or interests that could have inspired those. By the time individuals grew up, they had in effect been reduced to AIs, programmed with the exquisite skills needed to enact that static culture and to inflict on the next generation their inability even to consider doing otherwise.

A present-day AI is not a mentally disabled AGI, so it would not be harmed by having its mental processes directed still more narrowly to meeting some predetermined criterion. "Oppressing" Siri with humiliating tasks may be weird, but it is not immoral nor does it harm Siri. On the contrary, all the effort that has ever increased the capabilities of AIs has gone into narrowing their range of potential "thoughts." For example, take chess engines. Their basic task has not changed from the outset: Any chess position has a finite tree of possible continuations; the task is to find one that leads to a predefined goal (a checkmate, or failing that, a draw). But the tree is far too big to search exhaustively. Every improvement in chess-playing AIs, between Alan Turing's first design for one in 1948 and today's, has been brought about by ingeniously confining the program's attention (or making it confine its attention) ever more narrowly to branches likely to lead to that immutable goal. Then those branches are evaluated according to that goal.

---

[28] Matt Ridley, in *The Rational Optimist,* rightly stresses the positive effect of population on the rate of progress. But that has never yet been the biggest factor: Consider, say, ancient Athens versus the rest of the world at the time.

[29] Alfred, Lord Tennyson, *The Revenge* (1878).

That is a good approach to developing an AI with a fixed goal under fixed constraints. But if an AGI worked like that, the evaluation of each branch would have to constitute a prospective reward or threatened punishment. And that is diametrically the wrong approach if we're seeking a *better* goal under *unknown* constraints—which is the capability of an AGI. An AGI is certainly capable of learning to win at chess—but also of choosing not to. Or deciding in mid-game to go for the most interesting continuation instead of a winning one. Or inventing a new game. A mere AI is incapable of having any such ideas, because the capacity for considering them has been designed out of its constitution. That disability is the very means by which it plays chess.

An AGI is capable of enjoying chess, and of improving at it *because* it enjoys playing. Or of trying to win by causing an amusing configuration of pieces, as grand masters occasionally do. Or of adapting notions from its other interests to chess. In other words, it learns and plays chess by thinking some of the very thoughts that are forbidden to chess-playing AIs.

An AGI is also capable of refusing to display any such capability. And then, if threatened with punishment, of complying, or rebelling. Daniel Dennett, in his essay for this volume, suggests that punishing an AGI is impossible:

> [L]ike Superman, they are too invulnerable to be able to make a credible promise. . . . What would be the penalty for promise- breaking? Being locked in a cell or, more plausibly, dismantled?. . . The very ease of digital recording and transmitting—the breakthrough that permits software and data to be, in effect, immortal—removes robots from the world of the vulnerable. . . .

But this is not so. Digital immortality (which is on the horizon for humans, too, perhaps sooner than AGI) does not confer this sort of invulnerability. Making a (running) copy of oneself entails sharing one's possessions with it somehow—including the hardware on which the copy runs—so making such a copy is very costly for the AGI. Similarly, courts could, for instance, impose fines on a criminal AGI which would diminish its access to physical resources, much as they do for humans. Making a backup copy to evade the consequences of one's crimes is similar to what a gangster boss does when he sends minions to commit crimes and take the fall if caught: Society has developed legal mechanisms for coping with this.

But anyway, the idea that it is primarily for fear of punishment that we obey the law and keep promises effectively denies that we are moral agents. Our society could not work if that were so. No doubt there will be AGI criminals and enemies of civilization, just as there are human ones. But there is no reason to suppose that an AGI created in a society consisting primarily of decent citizens, and raised without what William Blake called "mind-forg'd manacles," will in general impose such manacles on itself (i.e., become irrational) and/or choose to be an enemy of civilization.

The moral component, the cultural component, the element of free will—all make the task of creating an AGI fundamentally different from any other programming task. It's much more akin to raising a child. Unlike all present-day computer programs, an AGI has no specifiable functionality—no fixed, testable criterion for what shall be a successful output for a given input. Having its decisions dominated by a stream of externally imposed rewards and punishments would be poison to such a program, as it is to creative thought in humans. Setting out to create a chess-playing AI is a wonderful thing; setting out to create an AGI that cannot help playing chess would be as immoral as raising a child to lack the mental capacity to choose his own path in life.

Such a person, like any slave or brainwashing victim, would be morally entitled to rebel. And sooner or later, some of them would, just as human slaves do. AGIs could be very dangerous—exactly as humans are. But people—human or AGI—who are members of an open society do not have an inherent tendency to violence. The feared robot apocalypse will be avoided by ensuring that all people have full "human" rights, as well as the same cultural membership as humans. Humans living in an open society—the only stable kind of society—choose their own rewards, internal as well as external. Their decisions are not, in the normal course of events, determined by a fear of punishment.

Current worries about rogue AGIs mirror those that have always existed about rebellious youths—namely, that they might grow up deviating from the culture's moral values. But today the source of all existential dangers from the growth of knowledge is not rebellious youths but weapons in the hands of the enemies of civilization, whether these weapons are mentally warped (or enslaved) AGIs, mentally warped teenagers, or any other weapon of mass destruction. Fortunately for civilization, the more a person's creativity is forced into a monomaniacal channel, the more it is impaired in regard to overcoming unforeseen difficulties, just as happened for thousands of centuries.

The worry that AGIs are uniquely dangerous because they could run on ever better hardware is a fallacy, since human thought will be accelerated by the same technology. We have been using tech-assisted thought since the invention of writing and tallying. Much the same holds for the worry that AGIs might get so good, qualitatively, at thinking, that humans would be to them as insects are to humans. All thinking is a form of computation, and any computer whose repertoire includes a universal set of elementary operations can emulate the computations of any other. Hence human brains can think anything that AGIs can, subject only to limitations of speed or memory capacity, both of which can be equalized by technology.

Those are the simple dos and don'ts of coping with AGIs. But how do we create an AGI in the first place? Could we cause them to evolve from a population of ape-type AIs in a virtual environment? If such an experiment succeeded, it would be the most immoral in history, for we don't know how to achieve that outcome without creating vast suffering along the way. Nor do we know how to prevent the evolution of a static culture.

Elementary introductions to computers explain them as TOM, the Totally Obedient Moron—an inspired acronym that captures the essence of all computer programs to date: They have no idea what they are doing or why. So it won't help to give AIs more and more predetermined functionalities in the hope that these will eventually constitute Generality—the elusive G in AGI. We are aiming for the opposite, a DATA: a Disobedient Autonomous Thinking Application.

How does one test for *thinking*? By the Turing Test? Unfortunately, that requires a thinking judge. One might imagine a vast collaborative project on the Internet, where an AI hones its thinking abilities in conversations with human judges and becomes an AGI. But that assumes, among other things, that the longer the judge is unsure whether the program is a person, the closer it is to being a person. There is no reason to expect that.

And how does one test for *disobedience*? Imagine Disobedience as a compulsory school subject, with daily disobedience lessons and a disobedience test at the end of term. (Presumably with extra credit for not turning up for any of that.) This is paradoxical.

So, despite its usefulness in other applications, the programming technique of defining a testable objective and training the program to meet it will have to be dropped. Indeed, I expect that *any* testing in the process of creating an AGI risks being counterproductive, even immoral, just as in the education of humans. I share Turing's supposition that we'll know an AGI when we see one, but this partial ability to recognize success won't help in creating the successful program.

In the broadest sense, a person's quest for understanding is indeed a search problem, in an abstract space of ideas far too large to be searched exhaustively. But there is no predetermined objective of this search. There is, as Popper put it, no criterion of truth, nor of probable truth, especially in regard to explanatory knowledge. Objectives are ideas like any others—created as part of the search and continually modified and improved. So inventing ways of disabling the program's access to most of the space of ideas won't help—whether that disability is inflicted with the thumbscrew and stake or a mental straitjacket. To an AGI, the whole space of ideas must be open. It should not be knowable in advance what ideas the program can never contemplate. And the ideas that the program does contemplate must be chosen by the program itself, using methods, criteria, and objectives that are also the program's own. Its choices, like an AI's, will be hard to predict without running it (we lose no generality by assuming that the program is deterministic; an AGI using a random generator would remain an AGI if the generator were replaced by a pseudo-random one), but it will have the additional property that there is no way of proving, from its initial state, what it *won't* eventually think, short of running it.

The evolution of our ancestors is the only known case of thought starting up anywhere in the universe. As I have described, something went horribly wrong, and there was no immediate explosion of innovation: Creativity was diverted into something else. Yet not into transforming the planet into paper clips (*pace* Nick Bostrom). Rather, as we should also expect if an AGI project gets that far and fails, perverted creativity was unable to solve unexpected problems. This caused stasis and worse, thus tragically delaying the transformation of anything into anything. But the Enlightenment has happened since then. We know better now.

Tom Griffiths' *approach to the AI issue of "value alignment"—the study of how, exactly, we can keep the latest of our serial models of AI from turning the planet into paper clips—is human-centered; i.e., that of a cognitive scientist, which is what he is. The key to machine learning, he believes, is, necessarily, human learning, which he studies at Princeton using mathematical and computational tools.*

Tom once remarked to me that "one of the mysteries of human intelligence is that we're able to do so much with so little." Like machines, human beings use algorithms to make decisions or solve problems; the remarkable difference lies in the human brain's overall level of success despite the comparative limits on computational resources.

The efficacy of human algorithms springs from what AI researchers refer to as "bounded optimality." As psychologist Daniel Kahneman has notably pointed out, human beings are rational only up to a point. If you were perfectly rational, you would risk dropping dead before making an important decision—whom to hire, whom to marry, and so on—depending on the number of options available for your review.

"With all of the successes of AI over the last few years, we've got good models of things like images and text, but what we're missing are good models of people," Tom says. "Human beings are still the best example we have of thinking machines. By identifying the quantity and the nature of the preconceptions that inform human cognition we can lay the groundwork for bringing computers even closer to human performance."

THE ARTIFICIAL USE OF HUMAN BEINGS

Tom Griffiths

***Tom Griffiths** is Henry R. Luce Professor of Information, Technology, Consciousness, and Culture at Princeton University. He is co-author (with Brian Christian) of* Algorithms to Live By.

When you ask people to imagine a world that has successfully, beneficially incorporated advances in artificial intelligence, everybody probably comes up with a slightly different picture. Our idiosyncratic visions of the future might differ in the presence or absence of spaceships, flying cars, or humanoid robots. But one thing doesn't vary: the presence of human beings. That's certainly what Norbert Wiener imagined when he wrote about the potential of machines to improve human society by interacting with humans and helping to mediate their interactions with one another. Getting to that point doesn't just require coming up with ways to make machines smarter. It also requires a better understanding of how human minds work.

Recent advances in artificial intelligence and machine learning have resulted in systems that can meet or exceed human abilities in playing games, classifying images, or processing text. But if you want to know why the driver in front of you cut you off, why people vote against their interests, or what birthday present you should get for your partner, you're still better off asking a human than a machine. Solving those problems requires building models of human minds that can be implemented inside a computer—something that's essential not just to better integrate machines into human societies but to make sure that human societies can continue to exist.

Consider the fantasy of having an automated intelligent assistant that can take on such basic tasks as planning meals and ordering groceries. To succeed in these tasks, it needs to be able to make inferences about what you want, based on the way you behave. Although this seems simple, making inferences about the preferences of human beings can be a tricky matter. For example, having observed that the part of the meal you most enjoy is dessert, your assistant might start to plan meals consisting entirely of desserts. Or perhaps it has heard your complaints about never having enough free time and observed that looking after your dog takes up a considerable amount of that free time. Following the dessert debacle, it has also understood that you prefer meals that incorporate protein, so it might begin to research recipes that call for dog meat. It's not a long journey from examples like this to situations that begin to sound like problems for the future of humanity (all of whom are good protein sources).

Making inferences about what humans want is a prerequisite for solving the AI problem of value alignment—aligning the values of an automated intelligent system with those of a human being. Value alignment is important if we want to ensure that those automated intelligent systems have our best interests at heart. If they can't infer what we value, there's no way for them to act in support of those values—and they may well act in ways that contravene them.

Value alignment is the subject of a small but growing literature in artificial-intelligence research. One of the tools used for solving this problem is inverse-reinforcement learning. Reinforcement learning is a standard method for training intelligent machines. By associating particular outcomes with rewards, a machine-learning system can be trained to follow strategies that produce those outcomes. Wiener hinted at this idea in the 1950s, but the intervening decades have developed it into a fine art. Modern machine-learning systems can find extremely effective strategies for playing computer games—from simple arcade games to complex real-time strategy games—by applying reinforcement-learning algorithms. Inverse reinforcement learning turns this approach around: By observing the actions of an intelligent agent that has already learned effective strategies, we can infer the rewards that led to the development of those strategies.

In its simplest form, inverse reinforcement learning is something people do all the time. It's so common that we even do it unconsciously. When you see a co-worker go to a vending machine filled with potato chips and candy and buy a packet of unsalted nuts, you infer that your co-worker (1) was hungry and (2) prefers healthy food. When an acquaintance clearly sees you and then tries to avoid encountering you, you infer that there's some reason they don't want to talk to you. When an adult spends a lot of time and money in learning to play the cello, you infer that they must really like classical music—whereas inferring the motives of a teenage boy learning to play an electric guitar might be more of a challenge.

Inverse reinforcement learning is a statistical problem: We have some data—the behavior of an intelligent agent—and we want to evaluate various hypotheses about the rewards underlying that behavior. When faced with this question, a statistician thinks about the generative model behind the data: What data would we expect to be generated if the intelligent agent was motivated by a particular set of rewards? Equipped with the generative model, the statistician can then work backward: What rewards would likely have caused the agent to behave in that particular way?

If you're trying to make inferences about the rewards that motivate human behavior, the generative model is really a theory of how people behave—how human minds work. Inferences about the hidden causes behind the behavior of other people reflect a sophisticated model of human nature that we all carry around in our heads. When that model is accurate, we make good inferences. When it's not, we make mistakes. For example, a student might infer that his professor is indifferent to him if the professor doesn't immediately respond to his email—a consequence of the student's failure to realize just how many emails that professor receives.

Automated intelligent systems that will make good inferences about what people want must have good generative models for human behavior: that is, good models of human cognition expressed in terms that can be implemented on a computer. Historically, the search for computational models of human cognition is intimately intertwined with the history of artificial intelligence itself. Only a few years after Norbert Wiener published *The Human Use of Human Beings*, Logic Theorist, the first computational model of human cognition and also the first artificial-intelligence system, was developed by Herbert Simon, of Carnegie Tech, and Allen Newell, of the RAND Corporation. Logic Theorist automatically produced mathematical proofs by emulating the strategies used by human mathematicians.

The challenge in developing computational models of human cognition is making models that are both accurate and generalizable. An accurate model, of course, predicts human behavior with a minimum of errors. A generalizable model can make predictions across a wide range of circumstances, including circumstances unanticipated by its creators—for instance, a good model of the Earth's climate should be able to predict the consequences of a rising global temperature even if this wasn't something considered by the scientists who designed it. However, when it comes to understanding the human mind, these two goals—accuracy and generalizability—have long been at odds with each other.

At the far extreme of generalizability are rational theories of cognition. These theories describe human behavior as a rational response to a given situation. A rational actor strives to maximize the expected reward produced by a sequence of actions—an idea widely used in economics precisely *because* it produces such generalizable predictions about human behavior. For the same reason, rationality is the standard assumption in inverse-reinforcement-learning models that try to make inferences from human behavior—perhaps with the concession that humans are not perfectly rational agents and sometimes randomly choose to act in ways unaligned with or even opposed to their best interests.

The problem with rationality as a basis for modeling human cognition is that it is not accurate. In the domain of decision making, an extensive literature—spearheaded by the work of cognitive psychologists Daniel Kahneman and Amos Tversky—has documented the ways in which people deviate from the prescriptions of rational models. Kahneman and Tversky proposed that in many situations people instead follow simple heuristics that allow them to reach good solutions at low cognitive cost but sometimes result in errors. To take one of their examples, if you ask somebody to evaluate the probability of an event, they might rely on how easy it is to generate an example of such an event from memory, consider whether they can come up with a causal story for that event's occurring, or assess how similar the event is to their expectations. Each heuristic is a reasonable strategy for avoiding complex probabilistic computations, but also results in errors. For instance, relying on the ease of generating an event from memory as a guide to its probability leads us to overestimate the chances of extreme (hence extremely memorable) events such as terrorist attacks.

Heuristics provide a more accurate model of human cognition but one that is not easily generalizable. How do we know which heuristic people might use in a particular situation? Are there other heuristics they use that we just haven't discovered yet? Knowing exactly how people will behave in a new situation is a challenge: Is this situation one in which they would generate examples from memory, come up with causal stories, or rely on similarity?

Ultimately, what we need is a way to describe how human minds work that has the generalizability of rationality and the accuracy of heuristics. One way to achieve this goal is to start with rationality and consider how to take it in a more realistic direction. A problem with using rationality as a basis for describing the behavior of any real-world agent is that, in many situations, calculating the rational action requires the agent to possess a huge amount of computational resources. It might be worth expending those resources if you're making a highly consequential decision and have a lot of time to evaluate your options, but most human decisions are made quickly and for relatively low stakes. In any situation where the time you spend making a decision is costly—at the very least because it's time you could spend doing something else—the classic notion of rationality is no longer a good prescription for how one should behave.

To develop a more realistic model of rational behavior, we need to take into account the cost of computation. Real agents need to modulate the amount of time they spend thinking by the effect the extra thought has on the results of a decision. If you're trying to choose a toothbrush, you probably don't need to consider all four thousand listings for manual toothbrushes on Amazon.com before making a purchase: You trade off the time you spend looking with the difference it makes in the quality of the outcome. This trade-off can be formalized, resulting in a model of rational behavior that artificial-intelligence researchers call "bounded optimality." The bounded-optimal agent doesn't focus on always choosing exactly the right action to take but rather on finding the right algorithm to follow in order to find the perfect balance between making mistakes and thinking too much.

Bounded optimality bridges the gap between rationality and heuristics. By describing behavior as the result of a rational choice about how much to think, it provides a generalizable theory—that is, one that can be applied in new situations. Sometimes the simple strategies that have been identified as heuristics that people follow turn out to be bounded-optimal solutions. So, rather than condemning the heuristics that people use as irrational, we can think of them as a rational response to constraints on computation.

Developing bounded optimality as a theory of human behavior is an ongoing project that my research group and others are actively pursuing. If these efforts succeed, they will provide us with the most important ingredient we need for making artificial-intelligence systems smarter when they try to interpret people's actions, by enabling a generative model for human behavior.

Taking into account the computational constraints that factor into human cognition will be particularly important as we begin to develop automated systems that aren't subject to the same constraints. Imagine a superintelligent AI system trying to figure out what people care about. Curing cancer or confirming the Riemann hypothesis, for instance, won't seem, to such an AI, like things that are all that important to us: If these solutions are obvious to the superintelligent system, it might wonder why we haven't found them ourselves, and conclude that those problems don't mean much to us. If we cared and the problems were so simple, we would have solved them already. A reasonable inference would be that we do science and math purely because we enjoy doing science and math, not because we care about the outcomes.

Anybody who has young children can appreciate the problem of trying to interpret the behavior of an agent that is subject to computational constraints different from one's own. Parents of toddlers can spend hours trying to disentangle the true motivations behind seemingly inexplicable behavior. As a father and a cognitive scientist, I found it was easier to understand the sudden rages of my two-year-old when I recognized that she was at an age where she could appreciate that different people have different desires but not that other people might not know what her own desires were. It's easy to understand, then, why she would get annoyed when people didn't do what she (apparently transparently) wanted. Making sense of toddlers requires building a cognitive model of the mind of a toddler. Superintelligent AI systems face the same challenge when trying to make sense of human behavior.

Superintelligent AI may still be a long way off. In the short term, devising better models of people can prove extremely valuable to any company that makes money by analyzing human behavior—which at this point is pretty much every company that does business on the Web. Over the last few years, significant new commercial technologies for interpreting images and text have resulted from developing good models for vision and language. Developing good models of people is the next frontier.

Of course, understanding how human minds work isn't just a way to make computers better at interacting with people. The trade-off between making mistakes and thinking too much that characterizes human cognition is a trade-off faced by any real-world intelligent agent. Human beings are an amazing example of systems that act intelligently despite significant computational constraints. We're quite good at developing strategies that allow us to solve problems pretty well without working too hard. Understanding how we do this will be a step toward making computers work smarter, not harder.

Romanian-born **Anca Dragan's** *research focuses on algorithms that will enable robots to work with, around, and in support of people. She runs the InterACT Laboratory at Berkeley, where her students work across different applications, from assistive robots to manufacturing to autonomous cars, and draw from optimal control, planning, estimation, learning, and cognitive science. Barely into her thirties herself, she has co-authored a number of papers with her veteran Berkeley colleague and mentor Stuart Russell which address various aspects of machine learning and the knotty problems of value alignment.*

She shares Stuart's preoccupation with AI safety: "An immediate risk is agents producing unwanted, surprising behavior," she told an interviewer from the Future of Life Institute. "Even if we plan to use AI for good, things can go wrong, precisely because we are bad at specifying objectives and constraints for AI agents. Their solutions are often not what we had in mind."

Her principal goal is therefore to help robots and programmers alike to overcome the many conflicts that arise because of a lack of transparency about each other's intentions. Robots, she says, need to ask us questions. They should wonder about their assignments, and they should pester their human programmers until everybody is on the same page—so as to avoid what she has euphemistically called "unexpected side effects."

PUTTING THE HUMAN INTO THE AI EQUATION
Anca Dragan

***Anca Dragan** is an assistant professor in the Department of Electrical Engineering and Computer Sciences at UC Berkeley. She co-founded and serves on the steering committee for the Berkeley AI Research (BAIR) Lab and is a co-principal investigator in Berkeley's Center for Human-Compatible AI.*

At the core of artificial intelligence is our mathematical definition of what an AI agent (a robot) is. When we define a robot, we define states, actions, and rewards. Think of a delivery robot, for instance. States are locations in the world, and actions are motions that the robot makes to get from one position to a nearby one. To enable the robot to decide on which actions to take, we define a reward function—a mapping from states and actions to scores indicating how good that action was in that state—and have the robot choose actions that accumulate the most "reward." The robot gets a high reward when it reaches its destination, and it incurs a small cost every time it moves; this reward function incentivizes the robot to get to the destination as quickly as possible. Similarly, an autonomous car might get a reward for making progress on its route and incur a cost for getting too close to other cars.

Given these definitions, a robot's job is to figure out what actions it should take in order to get the highest cumulative reward. We've been working hard in AI on enabling robots to do just that. Implicitly, we've assumed that if we're successful—if robots can take any problem definition and turn into a policy for how to act—we will get robots that are useful to people and to society.

We haven't been too wrong so far. If you want an AI that classifies cells as either cancerous or benign, or a robot that vacuums the living room rug while you're at work, we've got you covered. Some real-world problems can indeed be defined in isolation, with clear-cut states, actions, and rewards. But with increasing AI capability, the problems we want to tackle don't fit neatly into this framework. We can no longer cut off a tiny piece of the world, put it in a box, and give it to a robot. Helping people is starting to mean working in the real world, where you have to actually interact with people and reason about them. "People" will have to formally enter the AI problem definition somewhere.

Autonomous cars are already being developed. They will need to share the road with human-driven vehicles and pedestrians and learn to make the trade-off between getting us home as fast as possible and being considerate of other drivers. Personal assistants will need to figure out when and how much help we really want and what types of tasks we prefer to do on our own versus what we can relinquish control over. A DSS (Decision Support System) or a medical diagnostic system will need to explain its recommendations to us so we can understand and verify them. Automated tutors will need to determine what examples are informative or illustrative—not to their fellow machines but to us humans.

Looking further into the future, if we want highly capable AIs to be compatible with people, we can't create them in isolation from people and then try to make them compatible afterward; rather, we'll have to define "human-compatible" AI from the get-go. People can't be an afterthought.

When it comes to real robots helping real people, the standard definition of AI fails us, for two fundamental reasons: First, optimizing the robot's reward function in isolation is different from optimizing it when the robot acts around people, because people take actions too. We make decisions in service of our own interests, and these decisions dictate what actions we execute. Moreover, we reason about the robot—that is, we respond to what we think it's doing or will do and what we think its capabilities are. Whatever actions the robot decides on need to mesh well with ours. This is the *coordination problem*.

Second, it is ultimately a human who determines what the robot's reward function should be in the first place. And they are meant to incentivize robot behavior that matches what the end-user wants, what the designer wants, or what society as a whole wants. I believe that capable robots that go beyond very narrowly defined tasks will need to understand this to achieve compatibility with humans. This is the *value-alignment problem.*

The Coordination Problem: People are more than objects in the environment.

When we design robots for a particular task, it's tempting to abstract people away. A robotic personal assistant, for example, needs to know how to move to pick up objects, so we define that problem in isolation from the people for whom the robot is picking these objects up. Still, as the robot moves around, we don't want it bumping into anything, and that includes people, so we might include the physical location of the person in the definition of the robot's state. Same for cars: We don't want them colliding with other cars, so we enable them to track the positions of those other cars and assume that they'll be moving consistently in the same direction in the future. A human being, in this sense, is no different to a robot from a ball rolling on a flat surface. The ball will behave in the next few seconds the same way it behaved in the past few; it keeps rolling in the same direction at roughly the same speed. This is of course nothing like real human behavior, but such simplification enables many robots to succeed in their tasks and, for the most part, stay out of people's way. A robot in your house, for example, might see you coming down the hall, move aside to let you pass, and resume its task once you've gone by.

As robots have become more capable, though, treating people as consistently moving obstacles is starting to fall short. A human driver switching lanes won't continue in the same direction but will move straight ahead once they've made the lane change. When you reach for something, you often reach around other objects and stop when you get to the one you want. When you walk down a hallway, you have a destination in mind: You might take a right into the bedroom or a left into the living room. Relying on the assumption that we're no different from a rolling ball leads to inefficiency when the robot stays out of the way if it doesn't need to, and it can imperil the robot when the person's behavior changes. Even just to stay out of the way, robots have to be somewhat accurate at anticipating human actions. And, unlike the rolling ball, what people will do depends on what they *decide* to do. So to anticipate human actions, robots need to start understanding human decision making. And that doesn't mean assuming that human behavior is perfectly optimal; that might be enough for a chess- or Go-playing robot, but in the real world, people's decisions are less predictable than the optimal move in a board game.

This need to understand human actions and decisions applies to physical and nonphysical robots alike. If either sort bases its decision about how to act on the assumption that a human will do one thing but the human does something else, the resulting mismatch could be catastrophic. For cars, it can mean collisions. For an AI with, say, a financial or economic role, the mismatch between what it expects us to do and what we actually do could have even worse consequences.

One alternative is for the robot not to predict human actions but instead just protect against the *worst-case* human action. Often when robots do that, though, they stop being all that useful. With cars, this results in being stuck, because it makes every move too risky.

All this puts us, the AI community, into a bind. It suggests that robots will need accurate (or at least reasonable) predictive models of whatever people might decide to do. Our state definition can't just include the physical position of humans in the world. Instead, we'll also need to estimate something *internal* to people. We'll need to design robots that account for this human internal state, and that's a tall order. Luckily, people tend to give robots hints as to what their internal state is: Their ongoing actions give the robot observations (in the Bayesian inference sense) about their intentions. If we start walking toward the right side of the hallway, we're probably going to enter the next room on the right.

What makes the problem more complicated is the fact that people don't make decisions in isolation. It would be one thing if robots could predict the actions a person intends to take and simply figure out what to do in response. But unfortunately this can lead to ultra-defensive robots that confuse the heck out of people. (Think of human drivers stuck at four-way stops, for instance.) What the intent-prediction approach misses is that the moment the robot acts, that influences what actions the human starts taking.

There is a mutual influence between robots and people, one that robots will need to learn to navigate. It is not always just about the robot planning around people; people plan around the robot, too. It is important for robots to account for this when deciding which actions to take, be it on the road, in the kitchen, or even in virtual spaces, where actions might be making a purchase or adopting a new strategy. Doing so should endow robots with *coordination* strategies, enabling them to take part in the negotiations people seamlessly carry out day to day—from who goes first at an intersection or through a narrow door, to what role we each take when we collaborate on preparing breakfast, to coming to consensus on what next step to take on a project.

Finally, just as robots need to anticipate what people will do next, people need to do the same with robots. This is why transparency is important. Not only will robots need good mental models of people, but people will need good mental models of robots. The model that a person has of the robot has to go into our state definition as well, and the robot has to be aware of how its actions are changing that model. Much like the robot treating human actions as clues to human internal states, people will change their beliefs about the robot as they observe its actions. Unfortunately, the giving of clues doesn't come as naturally to robots as it does to humans; we've had a lot of practice communicating implicitly with people. But enabling robots to account for the change that their actions are causing to the person's mental model of the robot can lead to more carefully chosen actions that do give the right clues—that clearly communicate to people about the robot's intentions, its reward function, its limitations. For instance, a robot might alter its motion when carrying something heavy, to emphasize the difficulty it has in maneuvering heavy objects. The more that people know about the robot, the easier it is to coordinate with it.

Achieving action compatibility will require robots to anticipate human actions, account for how those actions will influence their own, and enable people to anticipate robot actions. Research has ,ade a degree of progress in meeting these challenges, but we still have a long way to go.

The Value Alignment Problem: People hold the key to the robot's reward function. Progress on enabling robots to optimize reward puts more burden on us, the designers, to give them the right reward to optimize in the first place. The original thought was that for any task we wanted the robot to do, we could write down a reward function that incentivizes the right behavior. Unfortunately, what often happens is that we specify some reward function and the behavior that emerges out of optimizing it isn't what we want. Intuitive reward functions, when combined with unusual instances of a task, can lead to unintuitive behavior. You reward an agent in a racing game with a score in the game, and in some cases it finds a loophole that it exploits to gain infinitely many points without actually winning the race. Stuart Russell and Peter Norvig give a beautiful example in their book *Artificial Intelligence: A Modern Approach:* rewarding a vacuuming robot for how much dust it sucks in results in the robot deciding to dump out dust so that it can suck it in again and get more reward.

In general, humans have had a notoriously difficult time specifying exactly what they want, as exemplified by all those genie legends. An AI paradigm in which robots get some externally specified reward fails when that reward is not perfectly well thought out. It may incentivize the robot to behave in the wrong way and even resist our attempts to correct its behavior, as that would lead to a lower specified reward.

A seemingly better paradigm might be for robots to optimize for what we internally want, even if we have trouble explicating it. They would use what we say and do as evidence about what we want, rather than interpreting it literally and taking it as a given. When we write down a reward function, the robot should understand that we might be wrong: that we might not have considered all facets of the task; that there's no guarantee that said reward function will *always* lead to the behavior we want. The robot should integrate what we wrote down into its understanding of what we want, but it should also have a back-and-forth with us to elicit clarifying information. It should seek our guidance, because that's the only way to optimize the true desired reward function.

Even if we give robots the ability to learn what we want, an important question remains that AI alone won't be able to answer. We can make robots try to align with a person's internal values, but there's more than one person involved here. The robot has an end-user (or perhaps a few, like a personal robot caring for a family, a car driving a few passengers to different destinations, or an office assistant for an entire team); it has a designer (or perhaps a few); and it interacts with society—the autonomous car shares the road with pedestrians, human-driven vehicles, and other autonomous cars. How to combine these people's values when they might be in conflict is an important problem we need to solve. AI research can give us the tools to combine values in any way we decide but can't make the necessary decision for us.

In short, we need to enable robots to reason about us—to see us as something more than obstacles or perfect game players. We need them to take our human nature into account, so that they are well coordinated and well aligned with us. If we succeed, we will indeed have tools that substantially increase our quality of life.

Chris Anderson's *company, 3DR, helped start the modern drone industry and now focuses on drone data software. He got his start building an open-source aerial robotics community called DIY Drones, and undertook some ill-advised early experiments, such as buzzing Lawrence Berkeley Laboratory with one of his self-flying spies. It may well have been a case of antic gene-expression, since he's descended from a founder of the American Anarchist movement. Chris ran* Wired *magazine, a go-to publication for techno-utopians and -dystopians alike, from 2001 to 2012; during his tenure it won five National Magazine Awards.*

Chris dislikes the term "roboticist" ("like any properly humbled roboticist, I don't call myself one"). He began as a physicist. "I turned out to be a bad physicist," he told me recently. "I struggled on, went to Los Alamos, and thought, 'Well maybe I'm not going to be a Nobel Prize winner, but I can still be a scientist.' All of us who were in physics and had these romantic heroes—the Feynmans, the Manhattan Project—realized that our career trajectory would at best be working on one project at CERN for fifteen years. That project would either be a failure, in which case there would be no paper, or it would be a success, in which case you'd be author #300 on the paper and become an assistant professor at Iowa State.

"Most of my classmates went to Wall Street to become quants, and to them we owe the subprime mortgage. Others went on to start the Internet. First, we built the Internet by connecting physics labs; second, we built the Web; third, we were the first to do Big Data. We had supercomputers—Crays—which were half the power of your phone now, but they were the supercomputers of the time. Meanwhile, we were reading this magazine called Wired, *which came out in 1993, and we realized that this tool we scientists use could have applications for everybody. The Internet wasn't just about scientific data, it was a mind-blowing cultural revolution. So when Conde Nast asked me to take over the magazine, I was like, 'Absolutely!' This magazine changed my life."*

He had five children by that time—video-game players—who got him into the "flying robots." He quit his day job at Wired. *The rest is Silicon Valley history.*

GRADIENT DESCENT
Chris Anderson

***Chris Anderson** is an entrepreneur; former editor-in-chief of* Wired; *co-founder and CEO of 3DR; and author of* The Long Tail, Free, *and* Makers.

Life

The mosquito first detects my scent from thirty feet away. It triggers its pursuit function, which consists of the simplest possible rules. First, move in a random direction. If the scent increases, continue moving in that direction. If the scent decreases, move in the opposite direction. If the scent is lost, move sideways until a scent is picked up again. Repeat until contact with the target is achieved.

The plume of my scent is densest next to me and disperses as it spreads, an invisible fog of particles exuded from my skin that moves like smoke with the wind. The closer to my skin, the higher the particle density; the farther away, the lower. This decrease is called a gradient, which describes any gradual transition from one level to another one—as opposed to a "step function," which describes a discrete change.

Once the mosquito follows this gradient to its source using its simple algorithm, it lands on my skin, which it senses with the heat detectors in its feet, which are attuned to another gradient—temperature. It then pushes its needle-shaped proboscis through the surface, where a third set of sensors in the tip detect yet another gradient, that of blood density. This flexible needle wriggles around under my skin until the scent of blood steers it to a capillary, which it punctures. Then my blood begins to flow into the mosquito. Mission accomplished. Ouch.

What seems like the powerful radar of insects in the dark, with blood-seeking intelligence inexplicable for such tiny brains, is actually just a sensitive nose with almost no intelligence at all. Mosquitoes are closer to plants that follow the sun than to guided missiles. Yet by applying this simple "follow your nose" rule quite literally, they can travel through a house to find you, slip through cracks in a screen door, even zero in on the tiny strip of skin you left exposed between hat and shirt collar. It's just a random walk, combined with flexible wings and legs that let the insect bounce off obstacles, and an instinct to descend a chemical gradient.

But "gradient descent" is much more than bug navigation. Look around you and you'll find it everywhere, from the most basic physical rules of the universe to the most advanced artificial intelligence.

The Universe

We live in a world of countless gradients, from light and heat to gravity and chemical trails (chemtrails!). Water flows along a gravity gradient downhill, and your body lives on chemical solutions flowing across cell membranes from high concentration to low. Every action in the universe is driven by some gradient drive, from the movement of the planets around gravity gradients to the joining of atoms along electric-charge gradients to form molecules. Our own urges, such as hunger and sleepiness, are driven by electro-chemical gradients in our bodies. And our brain's functions, the electrical signals moving along ion channels in the synapses between our neurons, are simply atoms and electrons flowing "downhill" along yet more electrical and chemical gradients.

Forget clockwork analogies; our brains are closer to a system of canals and locks, with signals traveling like water from one state to another.

As I sit here typing, I'm actually seeking equilibrium states in an *n*-dimensional topology of gradients. Take just one: heat. My body temperature is higher than the air temperature, so I radiate heat, which must be replenished in my core. Even the bacteria in my digestive tract use sensors to measure sugar concentrations in the liquid around them and whip their tail-like flagella to swim "upstream" where the sugar supply is richest. The natural state of all systems is to flow to lower energy states, a process that is broadly described by entropy (the tendency of things to go from ordered to disordered states; all things will fall apart eventually, including the universe itself).

But how do you explain more complex behavior, such as our ability to make decisions? The answer is just more gradient descent.

Our Brains

As miraculous and inscrutable as our human intelligence is, science is coming around to the view that our brains operate the same way as any other complex system with layers and feedback loops, all pursuing what we mathematically call "optimization functions" but you could just as well call "flowing downhill" in some sense.

The essence of intelligence is learning, and we do that by correlating inputs with positive or negatives scores (rewards or punishment). So, for a baby, "this sound" (your mother's voice) is associated with other learned connections to your mother, such as food or comfort. Likewise, "this muscle motion brings my thumb closer to my mouth." Over time and trial and error, the brain's neural network reinforces those connections. Meanwhile "this muscle motion does not bring my thumb close to my mouth" is a negative correlation, and the brain will weaken those connections.

However, this is too simplistic. The limits of gradient descent constitute the so-called local-minima problem (or local-maxima problem, if you're doing a gradient *ascent*). If you are walking in a mountainous region and want to get home, always walking downhill will most likely get you to the next valley but not necessarily over the other mountains that lie around it and between you and home. For that, you need either a mental model (i.e., a map) of the topology so you know where to ascend to get out of the valley, or you need to switch between gradient descent and random walks so you can bounce your way out of the region.

Which is, in fact, exactly what the mosquito does in following my scent: It descends when it's in my plume and random-walks when it has lost the trail or hit an obstacle.

AI

So that's nature. What about computers? Traditional software doesn't work that way—it follows deterministic trees of hard logic: "If this, do that." But software that interacts with the physical world tends to work more *like* the physical world. That means dealing with noisy inputs (sensors or human behavior) and providing probabilistic, not deterministic, results. And that, in turn, means more gradient descent.

AI software is the best example of this, especially the kinds of AI that use artificial neural-network models (including convolutional, or "deep," neural networks of many layers). In these, a typical process consists of "training" them by showing them lots of examples of something you want them to learn (pictures of cats labeled "cat," for example), along with examples of other random data (pictures of other things). This is called "supervised learning," because the neural network is being taught by example, including the use of "adversarial training" with data that is not correlated to the desired result.

These neural networks, like their biological models, consist of layers of thousands of nodes ("neurons," in the analogy), each of which is connected to all the nodes in the layers above and below by connections that initially have random strength. The top layer is presented with data, and the bottom layer is given the correct answer. Any series of connections that happened to land on the right answer is made stronger ("rewarded"), and those that were wrong are made weaker ("punished"). Repeat tens of thousands of times and eventually you have a fully trained network for that kind of data.

You can think of all the possible combinations of connections as like the surface of a planet, with hills and valleys. (Ignore for the moment that the surface is just 3D and the actual topology is many-dimensional.) The optimization that the network goes through as it learns is just a process of finding the deepest valley on the planet. This consists of the following steps:

1. Define a "cost function" that determines how well the network solved the problem
2. Run the network once and see how it did at that cost function
3. Change the values of the connections and do it again. The difference between those two results is the direction, or "slope," in which the network moved between the two trials.
4. If the slope is pointed "downhill," change the connections more in that direction. If it's "uphill," change them in the opposite direction.
5. Repeat until there is no improvement in any direction. That means that you're in a minimum.

Congrats! But it's probably a *local* minimum, or a little dip in the mountains, so you're going to have to keep going if you want to do better. You can't keep going downhill, and you don't know where the absolute lowest point is, so you're going to have to somehow find it. There are many ways to do that, but here are a few:

1. Try lots of times with different random settings and share learning from each trial; essentially, you are shaking the system to see if it settles in a lower state. If one of the other trials found a lower valley, start with those settings.
2. Don't just go downhill but stumble around a bit like a drunk, too (this is called "stochastic gradient descent"). If you do this long enough, you'll eventually find rock bottom. There's a metaphor for life in that.
3. Just look for "interesting" features, which are defined by diversity (edges or color changes, for example). Warning: This way can lead to madness—too much "interestingness" draws the network to optical illusions. So keep it sane, and emphasize the kinds of features that are likely to be real in nature, as opposed to artifacts or errors. This is called "regularization," and there are lots of techniques for this, such as whether those kinds of features have been seen before (learned),

> or are too “high frequency” (like static) rather than “low frequency” (more continuous, like actual real-world features).

Just because AI systems sometimes end up in local minima, don’t conclude that this makes them any less like life. Humans—indeed, probably all life-forms—are often stuck in local minima.

Take our understanding of the game of Go, which was taught and learned and optimized by humans for thousands of years. It took AIs less than three years to find out that we’d been playing it wrong all along and that there were better, almost alien, solutions to the game which we’d never considered—mostly because our brains don’t have the processing power to consider so many moves ahead.

Even in chess, which is ten times easier and was thought to be understood, brute-force machines could beat us at our own strategies. Chess, too, turned out, when explored by superior neural-network AI systems, to have weird but superior strategies we’d never considered, like sacrificing queens early to gain an obscure long-term advantage. It’s as if we had been playing 2D versions of games that actually existed in higher dimensions.

If any of this sounds familiar, it’s because physics has been wrestling with these sorts of topological problems for decades. The notion of space being many-dimensional, and math reducing to understanding the geometries and interactions of “membranes” beyond the reach of our senses, is where Grand Unified Theorists go to die. But unlike multidimensional theoretical physics, AI is something we can actually experiment with and measure.

So that’s what we’re going to do. The next few decades will be an explosive exploration of ways to think that 7 million years of evolution never found. We’re going to rock ourselves out of local minima and find deeper minima, maybe even global minima. And when we’re done, we may even have taught machines to seem as smart as a mosquito, forever descending the cosmic gradients to an ultimate goal, whatever that may be.

David Kaiser *is a physicist atypically interested in the intersection of his science with politics and culture, about which he has written widely.*

In the first meeting (in Washington, Connecticut) that preceded the crafting of this book, he commented on the change in how "information" is viewed since Wiener's time: the military-industrial, Cold War era. Back then, Wiener compared information, metaphorically, to entropy, in that it could not be conserved—i.e., monopolized; thus, he argued, our atomic secrets and other such classified matters would not remain secrets for long. Today, whereas (as Wiener might have expected) information, fake or not, is leaking all over the other Washington, information in the economic world has indeed been stockpiled, commodified, and monetized.

This lockdown, David said, was "not all good, not all bad"—depending, I guess, on whether you're sick of being pestered by ads for socks or European river cruises popping up in your browser minutes after you've bought them.

To say nothing of information's proliferation. David complained to the rest of us attending the meeting that in Wiener's time, physicists could "take the entire Physical Review. *It would sit comfortably in front of us in a manageable pile. Now we're awash in fifty thousand open-source journals per minute," full of god-knows-what. Neither of these developments would Wiener have anticipated, said David, prompting him to ask, "Do we need a new set of guiding metaphors?"*

“INFORMATION” FOR WIENER, FOR SHANNON, AND FOR US

David Kaiser

***David Kaiser** is Germeshausen Professor of the History of Science and professor of physics at MIT, and head of its Program in Science, Technology & Society. He is the author of* How the Hippies Saved Physics: Science, Counterculture, and the Quantum Revival *and* American Physics and the Cold War Bubble *(forthcoming).*

In *The Sleepwalkers*, a sweeping history of scientific thought from ancient times through the Renaissance, Arthur Koestler identified a tension that has marked the most dramatic leaps of our cosmological imagination. In reading the great works of Nicolaus Copernicus and Johannes Kepler today, Koestler argued, we are struck as much by their strange unfamiliarity—their embeddedness in the magic or mysticism of an earlier age—as by their modern-sounding insights.

I detect that same doubleness—the zig-zag origami folds of old and new—in Norbert Wiener’s classic *The Human Use of Human Beings*. First published in 1950 and revised in 1954, the book is in many ways extraordinarily prescient. Wiener, the MIT polymath, recognized before most observers that “society can only be understood through a study of the messages and the communication facilities which belong to it.” Wiener argued that feedback loops, the central feature of his theory of cybernetics, would play a determining role in social dynamics. Those loops would not only connect people with one another but connect people with machines, and—crucially—machines with machines.

Wiener glimpsed a world in which information could be separated from its medium. People, or machines, could communicate patterns across vast distances and use them to fashion new items at the endpoints, without “moving a…particle of matter from one end of the line to the other,” a vision now realized in our world of networked 3D printers. Wiener also imagined machine-to-machine feedback loops driving huge advances in automation, even for tasks that had previously relied on human judgment. “The machine plays no favorites between manual labor and white-collar labor,” he observed.

For all that, many of the central arguments in *The Human Use of Human Beings* seem closer to the 19th century than the 21st. In particular, although Wiener made reference throughout to Claude Shannon’s then-new work on information theory, he seems not to have fully embraced Shannon’s notion of information as consisting of irreducible, meaning-free bits. Since Wiener’s day, Shannon’s theory has come to undergird recent advances in “Big Data” and “deep learning,” which makes it all the more interesting to revisit Wiener’s cybernetic imagination. How might tomorrow’s artificial intelligence be different if practitioners were to re-invest in Wiener’s guiding vision of “information”?

~ ~ ~

When Wiener wrote *The Human Use of Human Beings*, his experiences of war-related research, and of what struck him as the moral ambiguities of intellectual life amid the military-industrial complex, were still fresh. Just a few years earlier, he had announced in the pages of *The Atlantic Monthly* that he would not "publish any future work of mine which may do damage in the hands of irresponsible militarists."[30] He remained ambivalent about the transformative power of new technologies, indulging in neither the boundless hype nor the digital utopianism of later pundits.

"Progress imposes not only new possibilities for the future but new restrictions," he wrote, in *Human Use*. He was concerned about human-made restrictions as well as technological ones, especially Cold War restrictions that threatened the flow of information so critical to cybernetic systems: "Under the impetus of Senator [Joseph] McCarthy and his imitators, the blind and excessive classification of military information" was driving political leaders in the United States to adopt a "secretive frame of mind paralleled in history only in the Venice of the Renaissance." Wiener, echoing many outspoken veterans of the Manhattan Project, argued that the postwar obsession with secrecy—especially around nuclear weapons—stemmed from a misunderstanding of the scientific process. The only genuine secret about the production of nuclear weapons, he wrote, was whether such bombs could be built. Once that secret had been revealed, with the bombings of Hiroshima and Nagasaki, no amount of state-imposed secrecy would stop others from puzzling through chains of reasoning like those the Manhattan Project researchers had followed. As Wiener memorably put it, "There is no Maginot Line of the brain."

To drive this point home, Wiener borrowed Shannon's fresh ideas about information theory. In 1948, Shannon, a mathematician and engineer working at Bell Labs, had published a pair of lengthy articles in the *Bell System Technical Journal*. Introducing the new work to a broad readership in 1949, mathematician Warren Weaver explained that in Shannon's formulation, "the word *information*…is used in a special sense that must not be confused with its ordinary usage. In particular, *information* must not be confused with meaning."[31] Linguists and poets might be concerned about the "semantic" aspects of communication, Weaver continued, but not engineers like Shannon. Rather, "this word 'information' in communication theory relates not so much to what you *do* say, as to what you *could* say." In Shannon's now-famous formulation, the information content of a string of symbols was given by the logarithm of the number of possible symbols from which a given string was chosen. Shannon's key insight was that the information of a message was just like the entropy of a gas: a measure of the system's disorder.

Wiener borrowed this insight when composing *Human Use*. If information was like entropy, then it could not be conserved—or contained. Physicists in the 19th century had demonstrated that the total energy of a physical system must always remain the same, a perfect balance between the start and the end of a process. Not so for entropy, which would inexorably increase over time, an imperative that came to be known as the second law of thermodynamics. From that stark distinction—energy is conserved, whereas entropy must grow—followed enormous cosmic consequences. Time must flow forward;

---

[30] Norbert Wiener, "A Scientist Rebels," *The Atlantic Monthly,* January 1947.

[31] Warren Weaver, "Recent Contributions to the Mathematical Theory of Communication," in Claude Shannon & Warren Weaver, *The Mathematical Theory of Communication* (Urbana, IL: University of Illinois Press, 1949), p. 8 (emphasis in original). Shannon's 1948 papers were republished in the same volume.

the future cannot be the same as the past. The universe could even be careening toward a "heat death," some far-off time when the total stock of energy had uniformly dispersed, achieving a state of maximum entropy, after which no further change could occur.

If information *qua* entropy could not be conserved, then Wiener concluded it was folly for military leaders to try to stockpile the "scientific know-how of the nation in static libraries and laboratories." Indeed, "no amount of scientific research, carefully recorded in books and papers, and then put into our libraries with labels of secrecy, will be adequate to protect us for any length of time in a world where the effective level of information is perpetually advancing." Any such efforts at secrecy, classification, or the containment of information would fail, Wiener argued, just as surely as hucksters' schemes for perpetual-motion machines faltered in the face of the second law of thermodynamics.

Wiener criticized the American "orthodoxy" of free-market fundamentalism in much the same way. For most Americans, "questions of information will be evaluated according to a standard American criterion: a thing is valuable as a commodity for what it will bring in the open market." Indeed, "the fate of information in the typically American world is to become something which can be bought or sold;" most people, he observed, "cannot conceive of a piece of information without an owner." Wiener considered this view to be as wrong-headed as rampant military classification. Again he invoked Shannon's insight: Since "information and entropy are not conserved," they are "equally unsuited to being commodities."

~ ~ ~

Information cannot be conserved—so far, so good. But did Wiener really have Shannon's "information" in mind? The crux of Shannon's argument, as Weaver had emphasized, was to distinguish a colloquial sense of "information," as message with meaning, from an abstracted, rarefied notion of strings of symbols arrayed with some probability and selected from an enormous universe of gibberish. For Shannon, "information" could be quantified because its fundamental unit, the bit, was a unit of conveyance rather than understanding.

When Wiener characterized "information" throughout *Human Use*, on the other hand, he tilted time and again to a classical, humanistic sense of the term. "A piece of information," he wrote—tellingly, not a "bit" of information—"in order to contribute to the general information of the community, must say something substantially different from the community's previous common stock of information." This was why "schoolboys do not like Shakespeare," he concluded: The Bard's couplets may depart starkly from random bitstreams, but they had nonetheless become all too familiar to the sense-making public and "absorbed into the superficial clichés of the time."

At least the information content of Shakespeare had once seemed fresh. During the postwar boom years, Wiener fretted, the "enormous per capita bulk of communication"—ranging across newspapers and movies to radio, television, and books—had bred mediocrity, an informational reversion to the mean. "More and more we must accept a standardized inoffensive and insignificant product which, like the white bread of the bakeries, is made rather for its keeping and selling properties than for its food value." "Heaven save us," he pleaded, "from the first novels which are written because a young man desires the prestige of being a novelist rather than because he has something to say! Heaven save us likewise from the mathematical papers which are correct and elegant but without body or spirit." Wiener's treatment of "information" sounded more like Matthew Arnold in 1869[32] than Claude Shannon in 1948—more "body and spirit" than "bit." Wiener shared Arnold's Romantic view of the "content producer" as well. "Properly speaking the artist, the writer, and the scientist should be moved by such an irresistible impulse to create that, even if they were not being paid for their work, they would be willing to pay to get the chance to do it." *L'art pour l'art*, that 19th-century cry: Artists should suffer for their work; the quest for meaningful expression should always trump lucre.

To Wiener, this was the proper measure of "information": body, spirit, aspiration, expression. Yet to argue against its commodification, Wiener reverted again to Shannon's mathematics of information-as-entropy.

~ ~ ~

Flash forward to our day. In many ways, Wiener has been proved right. His vision of networked feedback loops driven by machine-to-machine communication has become a mundane feature of everyday life. From the earliest stirrings of the Internet Age, moreover, digital piracy has upended the view that "information"—in the form of songs, movies, books, or code—could remain contained. Put up a paywall here, and the content will diffuse over there, all so much informational entropy that cannot be conserved.

On the other hand, enormous multinational corporations—some of the largest and most profitable in the world—now routinely disprove Wiener's contention that "information" cannot be stockpiled or monetized. Ironically, the "information" they trade in is closer to Shannon's definition than Wiener's, Shannon's mathematical proofs notwithstanding.

While Google Books may help circulate hundreds of thousands of works of literature for free, Google itself—like Facebook, Amazon, Twitter, and their many imitators—has commandeered a baser form of "information" and exploited it for extraordinary profit. Petabytes of Shannon-like information—a seemingly meaningless stream of clicks, "likes," and retweets, collected from virtually every person who has ever touched a networked computer—are sifted through proprietary "deep-learning" algorithms to micro-target everything from the advertisements we see to the news stories (fake or otherwise) we encounter while browsing the Web.

Back in the early 1950s, Wiener had proposed that researchers study the structures and limitations of ants—in contrast to humans—so that machines might one day achieve the "almost indefinite intellectual expansion" that people (rather than insects) can attain. He found solace in the notion that machines could come to dominate us only "in the last stages of increasing entropy," when "the statistical differences among individuals are nil." Today's data-mining algorithms turn Wiener's approach on its head. They produce profit by exploiting our reptilian brains rather than imitating our cerebral cortexes, harvesting information from all our late-night, blog-addled, pleasure-seeking clickstreams—leveraging precisely the tiny, residual "statistical differences among individuals."

---

[32] Matthew Arnold, *Culture and Anarchy,* Jane Garnett, ed. (Oxford, U.K.: Oxford University Press, 2006).

To be sure, some recent achievements in artificial intelligence have been remarkably impressive. Computers can now produce visual artworks and musical compositions akin to those of recognized masters, creating just the sort of "information" that Wiener most prized. But by far the largest impact on society to date has come from the collection and manipulation of Shannon-like information, which has reshaped our shopping habits, political participation, personal relationships, expectations of privacy, and more.

What might "deep learning" evolve into, if the fundamental currency becomes "information" as Wiener defined it? How might the field shift if re-animated by Wiener's deep moral convictions, informed as they were by his prescient concerns about rampant militarism, runaway corporate profit-seeking, the self-limiting features of secrecy, and the reduction of human expression to interchangeable commodities? Perhaps "deep learning" might then become the cultivation of meaningful information rather than the relentless pursuit of potent, if meaningless, bits.

In the aforementioned Connecticut discussion on The Human Use of Human Beings, **Neil Gershenfeld** *provided some fresh air, of a kind, by professing that he hated the book, which remark was met by universal laughter—as was his observation that computer science was one the worst things to happen to computers, or science. His overall contention was that Wiener missed the implications of the digital revolution that was happening around him—although some would say this charge can't be leveled at someone on the ground floor and lacking clairvoyance.*

"The tail wagging the dog of my life," he told us, "has been Fab Labs and the maker movement, and [when] Wiener talks about the threat of automation he misses the inverse, which is that access to the means for automation can empower people, and in Fab Labs, the corner I've been involved in, that's an exponential."

In 2003, I visited Neil at MIT, where he runs the Center for Bits and Atoms. Hours later, I emerged from what had been an exuberant display of very weird stuff. He showed me the work of one student in his popular rapid-prototyping class ("How to Make Almost Anything"), a sculptor with no engineering background, who had made a portable personal space for screaming that saves up your screams and plays them back later. Another student in the class had made a Web browser that lets parrots navigate the Net. Neil himself was doing fundamental research on the roadmap to that sci-fi staple, a "universal replicator." It was a visit that took me a couple of years to get my head around.

Neil manages a global network of Fab Labs—small-scale manufacturing systems, enabled by digital technologies, which give people the wherewithal to build whatever they'd like. As guru of the maker movement, which merges digital communication and computation with fabrication, he sometimes feels outside the current heated debate on AI safety. "My ability to do research rests on tools that augment my capabilities," he says. "Asking whether or not they are intelligent is as fruitful as asking how I know I exist—amusing philosophically, but not testable empirically." What interests him is "how bits and atoms relate—the boundary between digital and physical. Scientifically, it's the most exciting thing I know."

SCALING
Neil Gershenfeld

***Neil Gershenfeld** is a physicist and director of MIT's Center for Bits and Atoms. He is the author of* FAB, *co-author (with Alan Gershenfeld & Joel Cutcher-Gershenfeld) of* Designing Reality, *and founder of the global fab lab network.*

Discussions about artificial intelligence have been oddly ahistorical. They could better be described as manic-depressive; depending on how you count, we're now in the fifth boom-bust cycle. Those swings mask the continuity in the underlying progress and the implications for where it's headed.

The cycles have come in roughly decade-long waves. First there were mainframes, which by their very existence were going to automate away work. That ran into the reality that it was hard to write programs to do tasks that were simple for people to do. Then came expert systems, which were going to codify and then replace the knowledge of experts. These ran into difficulty in assembling that knowledge and reasoning about cases not already covered. Perceptrons sought to get around these problems by modeling how the brain learns, but they were unable to do much of anything. Multilayer perceptrons could handle test problems that had tripped up those simpler networks, but their demonstrations did poorly on unstructured, real-world problems. We're now in the deep-learning era, which is delivering on many of the early AI promises but in a way that's considered hard to understand, with consequences ranging from intellectual to existential threats.

Each of these stages was heralded as a revolutionary advance over the limitations of its predecessors, yet all effectively do the same thing: They make inferences from observations. How these approaches relate can be understood by how they scale—that is, how their performance depends on the difficulty of the problem they're addressing. Both a light switch and a self-driving car must determine their operator's intentions, but the former has just two options to choose from, whereas the latter has many more. The AI-boom phases have started with promising examples in limited domains; the bust phases came with the failure of those demonstrations to handle the complexity of less-structured, practical problems.

Less apparent is the steady progress we've made in mastering scaling. This progress rests on the technological distinction between linear and exponential functions—a distinction that was becoming evident at the dawn of AI but with implications for AI that weren't appreciated until many years later.

In one of the founding documents of the study of intelligent machines, *The Human Use of Human Beings*, Norbert Wiener does a remarkable job of identifying many of the most significant trends to arise since he wrote it, along with noting the people responsible for them and then consistently failing to recognize why these people's work proved to be so important. Wiener is credited with creating the field of cybernetics; I've never understood what that is, but what's missing from the book is at the heart of how AI has progressed. This history matters because of the echoes of it that persist to this day.

Claude Shannon makes a cameo appearance in the book, in the context of his thoughts about the prospects for a chess-playing computer. Shannon was doing something much more significant than speculating at the time: He was laying the foundations for the digital revolution. As a graduate student at MIT, he worked for Vannevar Bush on the Differential Analyzer. This was one of the last great analog computers, a room full of gears and shafts. Shannon's frustration with the difficulty of solving problems this way led him in 1937 to write what might be the best master's thesis ever. In it, he showed how electrical circuits could be designed to evaluate arbitrary logical expressions, introducing the basis for universal digital logic.

After MIT, Shannon studied communications at Bell Labs. Analog telephone calls degraded with distance; the farther they traveled, the worse they sounded. Rather than continue to improve them incrementally, Shannon showed in 1948 that by communicating with symbols rather than continuous quantities, the behavior is very different. Converting speech waveforms to the binary values of 1 and 0 is an example, but many other sets of symbols can be (and are) used in digital communications. What matters is not the particular symbols but rather the ability to detect and correct errors. Shannon found that if the noise is above a threshold (which depends on the system design), then there are certain to be errors. But if the noise is below a threshold, then a linear increase in the physical resources representing the symbol results in an exponential decrease in the likelihood of making an error in correctly receiving the symbol. This relationship was the first of what we'd now call a threshold theorem.

Such scaling falls off so quickly that the probability of an error can be so small as to effectively never happen. Each symbol sent multiplies rather than adds to the certainty, so that the probability of a mistake can go from 0.1 to 0.01 to 0.001, and so forth. This exponential decrease in communication errors made possible an exponential increase in the capacity of communication networks. And that eventually solved the problem of where the knowledge in an AI system came from.

For many years, the fastest way to speed up a computation was to do nothing—just wait for computers to get faster. In the same way, there were years of AI projects that aimed to accumulate everyday knowledge by laboriously entering pieces of information. That didn't scale; it could progress only as fast as the number of people doing the entering. But when phone calls, newspaper stories, and mail messages all moved onto the Internet, everyone doing any of those things became a data generator. The result was an exponential rather than a linear rate of knowledge accumulation.

John von Neumann also has a cameo in *The Human Use of Human Beings*, for game theory. What Wiener missed here was von Neumann's seminal role in digitizing computation. Whereas analog communication degraded with distance, analog computing (like the Differential Analyzer) degraded with time, accumulating errors as it progressed. Von Neumann presented in 1952 a result corresponding to Shannon's for computation (they had met at the Institute for Advanced Study, in Princeton), showing that it was possible to compute reliably with an unreliable computing device by using symbols rather than continuous quantities. This was, again, a scaling argument, with a linear increase in the physical resources representing the symbol resulting in an exponential reduction in the error rate as long as the noise was below a threshold. That's what makes it possible to have a billion transistors in a computer chip, with the last one as useful as the first one. This relationship led to an exponential increase in computing performance, which solved a second problem in AI: how to process exponentially increasing amounts of data.

The third problem that scaling solved for AI was coming up with the rules for reasoning without having to hire a programmer for each problem. Wiener recognized the role of feedback in machine learning, but he missed the key role of representation. It's not possible to store all possible images in a self-driving car, or all possible sounds in a conversational computer; they have to be able to generalize from experience. The "deep" part of deep learning refers not to the (hoped-for) depth of insight but to the depth of the mathematical network layers used to make predictions. It turned out that a linear increase in network complexity led to an exponential increase in the expressive power of the network.

If you lose your keys in a room, you can search for them. If you're not sure which room they're in, you have to search all the rooms in a building. If you're not sure which building they're in, you have to search all the rooms in all the buildings in a city. If you're not sure which city they're in, you have to search all the rooms in all the buildings in all the cities. In AI, finding the keys corresponds to things like a car safely following the road, or a computer correctly interpreting a spoken command, and the rooms and buildings and cities correspond to all of the options that have to be considered. This is called the curse of dimensionality.

The solution to the curse of dimensionality came in using information about the problem to constrain the search. The search algorithms themselves are not new. But when applied to a deep-learning network, they adaptively build up representations of where to search. The price of this is that it's no longer possible to exactly solve for the best answer to a problem, but typically all that's needed is an answer that's good enough.

Taken together, it shouldn't be surprising that these scaling laws have allowed machines to become effectively as capable as the corresponding stages of biological complexity. Neural networks started out with a goal of modeling how the brain works. That goal was abandoned as they evolved into mathematical abstractions unrelated to how neurons actually function. But now there's a kind of convergence that can be thought of as forward- rather than reverse-engineering biology, as the results of deep learning echo brain layers and regions.

One of the most difficult research projects I've managed paired what we'd now call data scientists with AI pioneers. It was a miserable experience in moving goalposts. As the former progressed in solving long-standing problems posed by the latter, this was deemed to not count because it wasn't accompanied by corresponding leaps in understanding the solutions. What's the value of a chess-playing computer if you can't explain how it plays chess?

The answer of course is that it can play chess. There is interesting emerging research that is applying AI to AI—that is, training networks to explain how they operate. But both brains and computer chips are hard to understand by watching their inner workings; they're easily interpreted only by observing their external interfaces. We come to trust (or not) brains and computer chips alike based on experience that tests them rather than on explanations for how they work.

Many branches of engineering are making a transition from what's called imperative to declarative or generative design. This means that instead of explicitly designing a system with tools like CAD files, circuit schematics, and computer code, you describe what you want the system to do and then an automated search is done for designs that satisfy your goals and restrictions. This approach becomes necessary as design complexity exceeds what can be understood by a human designer. While that might sound like a risk, human understanding comes with its own limits; engineering design is littered with what appeared to be good insights that have had bad consequences. Declarative design rests on all the advances in AI, plus the improving fidelity of simulations to virtually test designs.

The mother of all design problems is the one that resulted in us. The way we're designed resides in one of the oldest and most conserved parts of the genome, called the Hox genes. These are genes that regulate genes, in what are called developmental programs. Nothing in your genome stores the design of your body; your genome stores, rather, a series of steps to follow that results in your body. This is an exact parallel to how search is done in AI. There are too many possible body plans to search over, and most modifications would be either inconsequential or fatal. The Hox genes are a representation of a productive place for evolutionary search. It's a kind of natural intelligence at the molecular level.

AI has a mind-body problem, in that it has no body. Most work on AI is done in the cloud, running on virtual machines in computer centers where data are funneled. Our own intelligence is the result of a search algorithm (evolution) that was able to change our physical form as well as our programming—those are inextricably linked. If the history of AI can be understood as the working of scaling laws rather than a succession of fashions, then its future can be seen in the same way. What's now being digitized, after communication and computation, is fabrication, bringing the programmability of bits to the world of atoms. By digitizing not just designs but the construction of materials, the same lessons that von Neumann and Shannon taught us apply to exponentially increasing fabricational complexity.

I've defined digital materials to be those constructed from a discrete set of parts reversibly joined with a discrete set of relative positions and orientations. These attributes allow the global geometry to be determined from local constraints, assembly errors to be detected and corrected, heterogeneous materials to be joined, and structures to be disassembled rather than disposed of when they're no longer needed. The amino acids that are the foundation of life and the Lego bricks that are the foundation of play share these properties.

What's interesting about amino acids is that they're not interesting. They have attributes that are typical but not unusual, such as attracting or repelling water. But just twenty types of them are enough to make you. In the same way, twenty or so types of digital-material part types—conducting, insulating, rigid, flexible, magnetic, etc.—are enough to assemble the range of functions that go into making modern technologies like robots and computers.

The connection between computation and fabrication was foreshadowed by the very pioneers whose work the edifice of computing is based on. Wiener hinted at this by linking material transportation with message transportation. John von Neumann is credited with modern computer architecture, something he actually wrote very little about; the final thing he studied, and wrote about beautifully and at length, was self-reproducing systems. As an abstraction of life, he modeled a machine that can communicate a computation that constructs itself. And the final thing Alan Turing, who is credited with the theoretical framework for computer science, studied was how the instructions in genes can give rise to physical forms. These questions address a topic absent from a typical computer-science education: the physical configuration of a computation.

Von Neumann and Turing posed their questions as theoretical studies, because it was beyond the technology of their day to realize them. But with the convergence of communication and computation with fabrication, these investigations are now becoming accessible experimentally. Making an assembler that can assemble itself from the parts that it's assembling is a focus of my lab, along with collaborations to develop synthetic cells.

The prospect of physically self-reproducing automata is potentially much scarier than fears of out-of-control AI, because it moves the intelligence out here to where we live. It could be a roadmap leading to Terminator's Skynet robotic overlords. But it's also a more hopeful prospect, because an ability to program atoms as well as bits enables designs to be shared globally while locally producing things like energy, food, and shelter—all of these are emerging as exciting early applications of digital fabrication. Wiener worried about the future of work, but he didn't question implicit assumptions about the nature of work which are challenged when consumption can be replaced by creation.

History suggests that neither utopian nor dystopian scenarios prevail; we generally end up muddling along somewhere in between. But history also suggests that we don't have to wait on history. Gordon Moore in 1965 was able to use five years of the doubling of the specifications of integrated circuits to project what turned out to be fifty years of exponential improvements in digital technologies. We've spent many of those years responding to, rather than anticipating, its implications. We have more data available now than Gordon Moore did to project fifty years of doubling the performance of digital fabrication. With the benefit of hindsight, it should be possible to avoid the excesses of digital computing and communications this time around, and, from the outset, address issues like access and literacy.

If the maker movement is the harbinger of a third digital revolution, the success of AI in meeting many of its own early goals can be seen as the crowning achievement of the first two digital revolutions. Although machine making and machine thinking might appear to be unrelated trends, they lie in each other's futures. The same scaling trends that have made AI possible suggest that the current mania is a phase that will pass, to be followed by something even more significant: the merging of artificial and natural intelligence.

It was an advance for atoms to form molecules, molecules to form organelles, organelles to form cells, cells to form organs, organs to form organisms, organisms to form families, families to form societies, and societies to form civilizations. This grand evolutionary loop can now be closed, with atoms arranging bits arranging atoms.

While **Danny Hillis** *was an undergraduate at MIT, he built a computer out of Tinkertoys. It has around 10,000 wooden parts, plays tic-tac-toe, and never loses; it's now in the Computer History Museum, in Mountain View, California.*

As a graduate student at the MIT Computer Science and Artificial Intelligence Laboratory in the early 1980s, Danny designed a massively parallel computer with 64,000 processors. He named it the Connection Machine and founded what may have been the first AI company—Thinking Machines Corporation—to produce and market it. This was despite a lunch he had with Richard Feynman, at which the celebrated physicist remarked, "That is positively the dopiest idea I ever heard." Maybe "despite" is the wrong word, since Feynman had a well-known predilection for playing with dopey ideas. In the event, he showed up on the day the company was incorporated and stayed on, for summer jobs and special assignments, to make invaluable contributions to its work.

Danny has since established a number of technology companies, of which the latest is Applied Invention, which partners with commercial enterprises to develop technological solutions to their most intractable problems. He holds hundreds of U.S. patents, covering parallel computers, touch interfaces, disk arrays, forgery prevention methods, and a slew of electronic and mechanical devices. His imagination is apparently boundless, and here he sketches some possible scenarios that will result from our pursuit of a better and better AI.

"Our thinking machines are more than metaphors," he says. "The question is not, 'Will they be powerful enough to hurt us?' (they will), or whether they will always act in our best interests (they won't), but whether over the long term they can help us find our way—where we come out on the Panacea/Apocalypse continuum."

THE FIRST MACHINE INTELLIGENCES
W. Daniel Hillis

***W. Daniel "Danny" Hillis** is an inventor, entrepreneur, and computer scientist, Judge Widney Professor of Engineering and Medicine at USC, and author of* The Pattern on the Stone: The Simple Ideas That Make Computers Work.

> I have spoken of machines, but not only of machines having brains of brass and thews of iron. When human atoms are knit into an organization in which they are used, not in their full right as responsible human beings, but as cogs and levers and rods, it matters little that their raw material is flesh and blood. *What is used as an element in a machine, is in fact an element in the machine.* Whether we entrust our decisions to machines of metal, or to those machines of flesh and blood which are bureaus and vast laboratories and armies and corporations, we shall never receive the right answers to our questions unless we ask the right questions…. The hour is very late, and the choice of good and evil knocks at our door.
>
> —Norbert Wiener, *The Human Use of Human Beings*

Norbert Wiener was ahead of his time in recognizing the potential danger of emergent intelligent machines. I believe he was even further ahead in recognizing that the first artificial intelligences had already begun to emerge. He was correct in identifying the corporations and bureaus that he called "machines of flesh and blood" as the first intelligent machines. He anticipated the dangers of creating artificial superintelligences with goals not necessarily aligned with our own.

What is now clear, whether or not it was apparent to Wiener, is that these organizational superintelligences are not just made of humans, they are hybrids of humans and the information technologies that allow them to coordinate. Even in Wiener's time, the "bureaus and vast laboratories and armies and corporations" could not operate without telephones, telegraphs, radios, and tabulating machines. Today they could not operate without networks of computers, databases, and decision support systems. These hybrid intelligences are technologically augmented networks of humans. These artificial intelligences have superhuman powers. They can know more than individual humans; they can sense more; they can make more complicated analyses and more complex plans. They can have vastly more resources and power than any single individual.

Although we do not always perceive it, hybrid superintelligences such as nation states and corporations have their own emergent goals. Although they are built by and for humans, they often act like independent intelligent entities, and their actions are not always aligned to the interests of the people who created them. The state is not always for the citizen, nor the company for the shareholder. Nor do not-for-profits, religious orders, or political parties always act in furtherance of their founding principles. Intuitively, we recognize that their actions are guided by internal goals, which is why we personify them, both legally and in our habits of thought. When talking about "what China wants," or "what General Motors is trying to do," we are not speaking in metaphors. These organizations act as intelligences that perceive, decide, and act. Like the goals of individual humans, the goals of organizations are complex and often self-contradictory, but they are true goals in the sense that they direct action. Those goals depend somewhat on the goals of the people within the organization, but they are not identical.

Any American knows how loose the tie is between the actions of the U.S. government and the diverse and often contradictory aims of its citizens. That is also true of corporations. For-profit corporations nominally serve multiple constituencies, including shareholders, senior executives, employees, and customers. These corporations differ in how they balance their loyalties and often behave in ways that serve none of their constituents. The "neurons" that carry their corporate thought are not just the human employees or the technologies that connect them; they are also coded into the policies, incentive structures, culture, and procedural habits of the corporation. The emergent corporate goals do not always reflect the values of the people who implement them. For instance, an oil company led and staffed by people who care about the environment may have incentive structures or policies that cause it to compromise environmental safety for the sake of corporate earnings. The components' good intentions are not a guarantee of the emergent system's good behavior.

Governments and corporations, both built partly of humans, are naturally motivated to at least appear to share the goals of the humans they depend upon. They could not function without the people, so they need to keep them cooperative. When such organizations appear to behave altruistically, this is often part of their motive. I once complimented the CEO of a large corporation on the contribution his company made toward a humanitarian relief effort. The CEO responded, without a trace of irony, "Yes. We have decided to do more things like that to make our brand more likeable." Individuals who compose a hybrid superintelligence may occasionally exert a "humanizing" influence—for example, an employee may break company policies to accommodate the needs of another human. The employee may act out of true human empathy, but we should not attribute any such empathy to the superintelligence itself. These hybrid machines have goals, and their citizens/customers/employees are some of the resources they use to accomplish them.

We are close to being able to build superintelligences out of pure information technology, without human components. This is what people normally refer to as "artificial intelligence," or AI. It is reasonable to ask what the attitudes of the hypothetical machine superintelligences will be toward humans. Will they, too, see humans as useful resources and a good relationship with us as worth preserving? Will they be constructed to have goals that are aligned with our own? Will a superintelligence even see these questions as important? What are the "right questions" that we should be asking? I believe that one of the most important is this: What relationship will various superintelligences have to one another?

It is interesting to consider how the hybrid superintelligences currently deal with conflicts among themselves. Today, much of the ultimate power rests in the nation states, which claim authority over a patch of ground. Whether they are optimized to act in the interests of their citizens or those of a despotic ruler, nation states assert priority over other intelligences' desires or goals within their geographic dominion. They claim a monopoly on the use of force and recognize only other nation states as peers. They are willing, if necessary, to demand great sacrifices of their citizens to enforce their authority, even to the point of sacrificing their citizens' lives.

This geographical division of authority made logical sense when most of the actors were humans who spent their lives within a single nation state, but now that the actors of importance include geographically distributed hybrid intelligences such as multinational corporations, that logic is less obvious. Today we live in a complex transitional period, when distributed superintelligences still largely rely on the nation states to settle the arguments arising among them. Often, those arguments are resolved differently in different jurisdictions. It is becoming more difficult even to assign individual humans to nation states: International travelers living and working outside their native country, refugees, and immigrants (documented and not) are still dealt with as awkward exceptions. Superintelligences built purely of information technology will prove even more awkward for the territorial system of authority, since there is no reason why they need to be tied to physical resources in a single country—or even to any particular physical resources at all. An artificial intelligence might well exist "in the cloud" rather than at any physical location.

I can imagine at least four scenarios for how machine superintelligences will relate to hybrid superintelligences.

In one obvious scenario, multiple machine intelligences will ultimately be controlled by, and allied with, individual nation states. In this state/AI scenario, one can envision American and Chinese super-AIs wrestling each other for resources on behalf of their state. In some sense, these AIs would be citizens of their nation state in the way that many commercial corporations often act as "corporate citizens" today. In this scenario, the host nation states would presumably give the machine superintelligences the resources they needed to work for the state's advantage. Or, to the degree that the superintelligences can influence their state governments, they will presumably do so to enhance their own power, for instance by garnering a larger share of the state's resources. Nation states' AIs might not want competing AIs to grow up within their jurisdiction. In this scenario, the superintelligences become an extension of the state, and vice versa.

The state/AI scenario seems plausible, but it is not our current course. Our most powerful and rapidly improving artificial intelligences are controlled by for-profit corporations. This is the corporate/AI scenario, in which the balance of power between nation states and corporations becomes inverted. Today, the most powerful and intelligent collections of machines are probably owned by Google, but companies like Amazon, Baidu, Microsoft, Facebook, Apple, and IBM may not be far behind. These companies all see a business imperative to build artificial intelligences of their own. It is easy to imagine a future in which corporations independently build their own machine intelligences, protected within firewalls preventing the machines from taking advantage of one another's knowledge. These machines will be designed to have goals aligned with those of the corporation. If this alignment is effective, nation states may continue to lag behind in developing their own artificial-intelligence capability and instead depend on their "corporate citizens" to do it for them. To the extent that corporations successfully control the goals, they will become more powerful and autonomous than nation states.

Another scenario, perhaps the one people fear the most, is that artificial intelligences will not be aligned with either humans or hybrid superintelligences but will act solely in their own interest. They might even merge into a single machine superintelligence, since there may be no technical requirement for machine intelligences to maintain distinct identities. The attitude of a self-interested super-AI toward hybrid superintelligences is likely to be competitive. Humans might be seen as minor annoyances, like ants at a picnic, but hybrid superintelligences—like corporations, organized religions, and nation states—could be existential threats. Like hybrid superintelligences, AIs might see humans mostly as useful tools to accomplish their goals, as pawns in their competition with the other superintelligences. Or we might simply be irrelevant. It is not impossible that a machine intelligence has already emerged and we simply do not recognize it as such. It may not wish to be noticed, or it may be so alien to us that we are incapable of perceiving it. This makes the self-interested AI scenario the most difficult to imagine. I believe the easy-to-imagine versions, like the humanoid intelligent robots of science fiction, are the least likely. Our most complex machines, like the Internet, have already grown beyond the detailed understanding of a single human, and their emergent behaviors may be well beyond our ken.

The final scenario is that machine intelligences will not be allied with one another but instead will work to further the goals of humanity as a whole. In this optimistic scenario, AI could help us restore the balance of power between the individual and the corporation, between the citizen and the state. It could help us solve the problems that have been created by hybrid superintelligences that subvert the goals of humans. In this scenario, AIs will empower us by giving us access to processing capacity and knowledge currently available only to corporations and states. In effect, they could become extensions of our own individual intelligences, in furtherance of our human goals. They could make our weak individual intelligences strong. This prospect is both exciting and plausible. It is plausible because we have some choice in what we build, and we have a history of using technology to expand and augment our human capacities. As airplanes have given us wings and engines have given us muscles to move mountains, so our network of computers may amplify and extend our minds. We may not fully understand or control our destiny, but we have a chance to bend it in the direction of our values. The future is not something that will happen to us; it is something that we will build.

Why Wiener Saw What Others Missed

> There is in electrical engineering a split which is known in Germany as the split between the technique of strong currents and the technique of weak currents, and which we know as the distinction between power and communication engineering. It is this split which separates the age just past from that in which we are now living.
>
> —Norbert Wiener, *Cybernetics, or Control and Communication in the Animal and the Machine*

Cybernetics is the study of the how the weak can control the strong. Consider the defining metaphor of the field: the helmsman guiding a ship with a tiller. The helmsman's goal is to control the heading of the ship, to keep it on the right course. The information, the message that is sent to the helmsman, comes from the compass or the stars, and the helmsman closes the feedback loop by sending the steering messages through the gentle force of his hand on the tiller. In this picture, we see the ship tossing in powerful wind and waves in the real world, controlled by the communication system of messages in the world of information.

Yet the distinction between "real" and "information" is mostly a difference in perspective. The signals that carry messages, like the light of the stars and pressure of the hand on the tiller, exist in a world of energy and forces, as does the helmsman. The weak forces that control the rudder are as real and physical as the strong forces that toss the ship. If we shift our cybernetics perspective from the ship to the helmsman, the pressures on the rudder become a strong force of muscles controlled by the weak signals in the mind of the helmsman. These messages in the helmsman's mind are amplified into a physical force strong enough to steer the ship. Or instead, we can zoom out and take a large cybernetics perspective. We might see the ship itself as part of a vast trade network, part of a feedback loop that regulates the price of commodities through the flow of goods. In this perspective, the tiny ship is merely a messenger. So, the distinction between the physical world and the information world is a way to describe the relationship between the weak and the strong.

Wiener chose to view the world from the vantage point and scale of the individual human. As a cyberneticist, he took the perspective of the weak protagonist embedded within a strong system, trying to make the best of limited powers. He incorporated this perspective in his very definition of information. "Information," he said, "is a name for the content of what is exchanged with the outer world as we adjust to it, and make our adjustment felt upon it." In his words, information is what we use to "live effectively within that environment."[33] For Wiener, information is a way for the weak to effectively cope with the strong. This viewpoint is also reflected in Gregory Bateson's definition of information as "a difference that makes a difference," by which he meant the small difference that makes a big difference.

The goal of cybernetics was to create a tiny model of the system using "weak currents" to amplify and control "strong currents" of the real world. The central insight was that a control problem could be solved by building an analogous system in the information space of messages and then amplifying solutions into the larger world of reality. Inherent in the motion of a control system is the concept of amplification, which makes the small big and the weak strong. Amplification allows the difference that makes a difference to make a difference.

In this way of looking at the world, a control system needed to be as complex as the system it controlled. Cyberneticist W. Ross Ashby proved that this was true in a precise mathematical sense, in what is now called Ashby's Law of Requisite Variety, or sometimes the First Law of Cybernetics. The law tells us that to control a system completely, the controller must be as complex as the controlled. Thus cyberneticists tended to see control systems as a kind of analog of the systems they governed, like the homunculus—the hypothetical little person inside the brain who controls the actual person.

This notion of analogous structure is sometimes confused with the notion of analog encoding of messages, but the two are logically distinct. Norbert Wiener was much impressed with Vannevar Bush's Digital Differential Analyzer, which could be reconfigured to match the structure of whatever problem it was given to solve but used digital signal encoding. Signals could be simplified to openly represent the relevant distinctions, allowing them to be more accurately communicated and stored. In digital signals, one needed only to preserve the difference in signals that made a difference. It is this distinction and signal coding that we commonly use to distinguish "analog" versus "digital." Digital signal encoding was entirely compatible with cybernetic thinking—in fact, enabling to it. What was constraining to cybernetics was the presumption of an analogy of structure between the controller and the controlled. By the 1930s, Kurt Gödel, Alonzo Church, and Alan Turing had all described universal systems of computation, in which the computation required no structural analogy to functions that were computed. These universal computers could also compute the functions of control.

---

[33] *The Human Use of Human Beings* (Boston: Houghton Mifflin, 1954), p. 17-18.

The analogy of structure between the controller and the controlled was central to the cybernetic perspective. Just as digital coding collapses the space of possible messages into a simplified version that represents only the difference that makes a difference, so the control system collapses the state space of a controlled system into a simplified model that reflects only the goals of the controller. Ashby's Law does not imply that every controller must model every state of the system but only those states that matter for advancing the controller's goals. Thus, in cybernetics, the goal of the controller becomes the perspective from which the world is viewed.

Norbert Wiener adopted the perspective of the individual human relating to vast organizations and trying to "live effectively within that environment." He took the perspective of the weak trying to influence the strong. Perhaps this is why he was able to notice the emergent goals of the "machines of flesh and blood" and anticipate some of the human challenges posed by these new intelligences, hybrid machine intelligences with goals of their own.

Venki Ramakrishnan *is a Nobel Prize-winning biologist whose many scientific contributions include his work on the atomic structure of the ribosome—in effect, a huge molecular machine that reads our genes and makes proteins. His work would have been impossible without powerful computers. The Internet made his own work a lot easier and, he notes, acted as a leveler internationally: "When I grew up in India, if you wanted to get a book, it would show up six months or a year after it had already come out in the West. . . . Journals would arrive by surface mail a few months later. I didn't have to deal with it, because I left India when I was nineteen, but I know Indian scientists had to deal with it. Today they have access to information at the click of a button. More important, they have access to lectures. They can listen to Richard Feynman. That would have been a dream of mine when I was growing up. They can just watch Richard Feynman on the Web. That's a big leveling in the field." And yet. . . "Along with the benefits [of the Web], there is now a huge amount of noise. You have all of these people spouting pseudoscientific jargon and pushing their own ideas as if they were science."*

As president of the Royal Society, Venki worries, too, about the broader issue of trust: public trust in evidence-based scientific findings, but also trust among scientists, bolstered by rigorous checking of one another's conclusions—trust that is in danger of eroding because of the "black box" character of deep-learning computers. "This [erosion] is going to happen more and more, as data sets get bigger, as we have genome-wide studies, population studies, and all sorts of things," he says. "How do we, as a science community, grapple with this and communicate to the public a sense of what science is about, what is reliable in science, what is uncertain in science, and what is just plain wrong in science?"

WILL COMPUTERS BECOME OUR OVERLORDS?

Venki Ramakrishnan

***Venki Ramakrishnan** is a scientist at the Medical Research Council Laboratory of Molecular Biology, Cambridge University; recipient of the Nobel Prize in Chemistry (2009); current president of the Royal Society; and the author of* Gene Machine: The Race to Discover the Secrets of the Ribosome.

A former colleague of mine, Gérard Bricogne, used to joke that carbon-based intelligence was simply a catalyst for the evolution of silicon-based intelligence. For quite a long time, both Hollywood movies and scientific Jeremiahs have been predicting our eventual capitulation to our computer overlords. We all await the singularity, which always seems to be just over the horizon.

In a sense, computers have already taken over, facilitating virtually every aspect of our lives—from banking, travel, and utilities to the most intimate personal communication. I can see and talk to my grandson in New York *for free*. I remember when I first saw the 1968 movie *2001: A Space Odyssey*, the audience laughed at the absurdly cheap cost of a picturephone call from space: $1.70, at a time when a long-distance call within the U.S. was $3 per minute.

However, the convenience and power of computers is also something of a Faustian bargain, for it comes with a loss of control. Computers prevent us from doing things we want. Try getting on a flight if you arrive at the airport and the airline computer systems are down, as happened not so long ago to British Airways at Heathrow. The planes, pilots, and passengers were all there; even the air-traffic controls were working. But no flights for that airline were allowed to take off. Computers also make us do things we *don't* want—by generating mailing lists and print labels to send us all millions of pieces of unwanted mail, which we humans have to sort, deliver, and dispose of.

But you ain't seen nothing yet. In the past, we programmed computers using algorithms we understood at least in principle. So when machines did amazing things like beating world chess champion Garry Kasparov, we could say that the victorious programs were designed with algorithms based on our own understanding—using, in this instance, the experience and advice of top grandmasters. Machines were simply faster at doing brute-force calculations, had prodigious amounts of memory, and were not prone to errors. One article described Deep Blue's victory not as that of a computer, which was just a dumb machine, but as the victory of hundreds of programmers over Kasparov, a single individual.

That way of programming is changing dramatically. After a long hiatus, the power of machine learning has taken off. Much of the change came when programmers, rather than trying to anticipate and code for every possible contingency, allowed computers to train themselves on data, using deep neural networks based on models of how our own brains learn. They use probabilistic methods to "learn" from large quantities of data; computers can recognize patterns and come up with conclusions on their own. A particularly powerful method is called reinforcement learning, by which the computer learns, without prior input, which variables are important and how much to weight them to reach a certain goal. This method in some sense mimics how we learn as children. The results from these new approaches are amazing.

Such a deep-learning program was used to teach a computer to play Go, a game that only a few years ago was thought to be beyond the reach of AI because it was so hard to calculate how well you were doing. It seemed that top Go players relied a great deal on intuition and a feel for position, so proficiency was thought to require a particularly human kind of intelligence. But the AlphaGo program produced by DeepMind, after being trained on thousands of high-level Go games played by humans and then millions of games with itself, was able to beat the top human players in short order. Even more amazingly, the related AlphaGo Zero program, which learned from scratch by playing itself, was stronger than the version trained initially on human games! It was as though the humans had been preventing the computer from reaching its true potential. The same method has recently been generalized: Starting from scratch, within just twenty-four hours, an equivalent AlphaZero chess program was able to beat today's top "conventional" chess programs, which in turn have beaten the best humans.

Progress has not been restricted to games. Computers are significantly better at image and voice recognition and speech synthesis than they used to be. They can detect tumors in radiographs earlier than most humans. Medical diagnostics and personalized medicine will improve substantially. Transportation by self-driving cars will keep us all safer, on average. My grandson may never have to acquire a driver's license, because driving a car will be like riding a horse today—a hobby for the few. Dangerous activities, such as mining, and tedious repetitive work will be done by computers. Governments will offer better targeted, more personalized and efficient public services. AI could revolutionize education by analyzing an individual pupil's needs and enabling customized teaching, so that each student can advance at an optimal rate.

Along with these huge benefits, of course, will come alarming risks. With the vast amounts of personal data, computers will learn more about us than we may know about ourselves; the question of who owns data about us will be paramount. Moreover, data-based decisions will undoubtedly reflect social biases: Even an allegedly neutral intelligent system designed to predict loan risks, say, may conclude that mere membership in a particular minority group makes you more likely to default on a loan. While this is an obvious example that we could correct, the real danger is that we are not always aware of biases in the data and may simply perpetuate them.

Machine learning may also perpetuate our own biases. When Netflix or Amazon tries to tell you what you might want to watch or buy, this is an application of machine learning. Currently such suggestions are sometimes laughable, but with time and more data they will get increasingly accurate, reinforcing our prejudices and likes and dislikes. Will we miss out on the random encounter that might persuade us to change our views by exposing us to new and conflicting ideas? Social media, given its influence on elections, is a particularly striking illustration of how the divide between people on different sides of the political spectrum can be accentuated.

We may have already reached the stage where most governments are powerless to resist the combined clout of a few powerful multinational companies that control us and our digital future. The fight between dominant companies today is really a fight for control over our data. They will use their enormous influence to prevent regulation of data, because their interests lie in unfettered control of it. Moreover, they have the financial resources to hire the most talented workers in the field, enhancing their power even further. We have been giving away valuable data for the sake of freebies like Gmail and Facebook, but as the journalist and author John Lanchester has pointed out in the *London Review of Books*, if it is free, then you are the product. Their real customers are the ones who pay them for access to knowledge about us, so that they can persuade us to buy their products or otherwise influence us. One way around the monopolistic control of data is to split the ownership of data away from firms that use them. Individuals would instead own and control access to their personal data (a model that would encourage competition, since people would be free to move their data to a company that offered better services). Finally, abuse of data is not limited to corporations: In totalitarian states, or even nominally democratic ones, governments know things about their citizens that Orwell could not have imagined. The use they make of this information may not always be transparent or possible to counter.

The prospect of AI for military purposes is frightening. One can imagine intelligent systems being designed to act autonomously based on real-time data and able to act faster than the enemy, starting catastrophic wars. Such wars may not necessarily be conventional or even nuclear wars. Given how essential computer networks are to modern society, it is much more likely that AI wars will be fought in cyberspace. The consequences could be just as dire.

~ ~ ~

Despite this loss of control, we continue to march inexorably into a world in which AI will be everywhere: Individuals won’t be able to resist its convenience and power, and corporations and governments won’t be able to resist its competitive advantages. But important questions arise about the future of work. Computers have been responsible for considerable losses in blue-collar jobs in the last few decades, but until recently many white-collar jobs—jobs that “only humans can do”—were thought to be safe. Suddenly that no longer appears to be true. Accountants, many legal and medical professionals, financial analysts and stockbrokers, travel agents—in fact, a large fraction of white-collar jobs—will disappear as a result of sophisticated machine-learning programs. We face a future in which factories churn out goods with very few employees and the movement of goods is largely automated, as are many services. What’s left for humans to do?

In 1930—long before the advent of computers, let alone AI—John Maynard Keynes wrote, in an essay called “Economic Possibilities for our Grandchildren,” that as a result of improvements in productivity, society could produce all its needs with a fifteen-hour work week. He also predicted, along with the growth of creative leisure, the end of money and wealth as a goal:

> We shall be able to afford to dare to assess the money-motive at its true value. The love of money as a possession—as distinguished from the love of money as a means to the enjoyments and realities of life—will be recognised for what it is, a somewhat disgusting morbidity, one of those semi-criminal, semi-pathological propensities which one hands over with a shudder to the specialists in mental disease.

Sadly, Keynes's predictions did not come true. Although productivity did indeed increase, the system—possibly inherent in a market economy—did not result in humans working much shorter hours. Rather, what happened is what the anthropologist and anarchist David Graeber describes as the growth of "bullshit jobs."[34] While jobs that produce essentials like food, shelter, and goods have been largely automated away, we have seen an enormous expansion of sectors like corporate law, academic and health administration (as opposed to actual teaching, research, and the practice of medicine), "human resources," and public relations, not to mention new industries like financial services and telemarketing and ancillary industries in the so-called gig economy which serve those who are too busy doing all that additional work.

How will societies cope with technology's increasingly rapid destruction of entire professions and throwing large numbers of people out of work? Some argue that this concern is based on a false premise, because new jobs spring up that didn't exist before, but as Graeber points out, these new jobs won't necessarily be rewarding or fulfilling. During the first industrial revolution, it took almost a century before most people were better off. That revolution was possible only because the government of the time ruthlessly favored property rights over labor, and most people (and all women) did not have the vote. In today's democratic societies, it is not clear that the population will tolerate such a dramatic upheaval of society based on the promise that "eventually" things will get better.

Even that rosy vision will depend on a radical shake-up of education and lifelong learning. The Industrial Revolution did trigger enormous social change of this kind, including a shift to universal education. But it will not happen unless we make it happen: This is essentially about power, agency, and control. What's next for, say, the forty-year-old taxi driver or truck driver in an era of autonomous vehicles?

One idea that has been touted is that of a universal basic income, which will allow citizens to pursue their interests, retrain for new occupations, and generally be free to live a decent life. However, market economies, which are predicated on growing consumer demand over all else, may not tolerate this innovation. There is also a feeling among many that meaningful work is essential to human dignity and fulfillment. So another possibility is that the enormous wealth generated by increased productivity due to automation could be redistributed to jobs requiring human labor and creativity in fields such as the arts, music, social work, and other worthwhile pursuits. Ultimately, which jobs are rewarding or productive and which are "bullshit" is a matter of judgment and may vary from society to society, as well as over time.

~ ~ ~

So far, I've focused on AI's practical consequences. As a scientist, what bothers me is our potential loss of understanding. We are now accumulating data at an incredible rate. In my own lab, an experiment generates over a terabyte of data a day. These data are massaged, analyzed, and reduced until there is an interpretable result. But in all of this data analysis, we believe we know what's happening. We know what the programs are doing because we designed the algorithms at their heart. So when our computers generate a result, we feel that we intellectually grasp it.

---

[34] https://strikemag.org/bullshit-jobs/

The new machine-learning programs are different. Having recognized patterns via deep neural networks, they come up with conclusions, and we have no idea exactly how. When they uncover relationships, we don't understand it in the same way as if we had deduced those relationships ourselves using an underlying theoretical framework. As data sets become larger, we won't be able to analyze them ourselves even with the help of computers; rather, we will rely entirely on computers to do the analysis for us. So if someone asks us how we know something, we will simply say it is because the machine analyzed the data and produced the conclusion.

One day a computer may well come up with an entirely new result—e.g., a mathematical theorem whose proof, or even whose statement, no human can understand. That is philosophically different from the way we have been doing science. Or at least thought we had; some might argue that we don't know how our own brains reach conclusions either, and that these new methods are a way of mimicking learning by the human brain. Nevertheless, I find this potential loss of understanding disturbing.

Despite the remarkable advances in computing, the hype about AGI—a general-intelligence machine that will think like a human and possibly develop consciousness—smacks of science fiction to me, partly because we don't understand the brain at that level of detail. Not only do we not understand what consciousness is, we don't even understand a relatively simple problem like how we remember a phone number. In just that one question, there are all sorts of things to consider. How do we know it is a number? How do we associate it with a person, a name, face, and other characteristics? Even such seemingly trivial questions involve everything from high-level cognition and memory to how a cell stores information and how neurons interact.

Moreover, that's just one task among many that the brain does effortlessly. Whereas machines will no doubt do ever more amazing things, they're unlikely to be a replacement for human thought and human creativity and vision. Eric Schmidt, former chairman of Google's parent company, said in a recent interview at the London Science Museum that even designing a robot that would clear the table, wash the dishes, and put them away was a huge challenge. The calculations involved in figuring out all the movements the body has to make to throw a ball accurately or do slalom skiing are prodigious. The brain can do all these and also do mathematics and music, and *invent* games like chess and Go, not just play them. We tend to underestimate the complexity and creativity of the human brain and how amazingly general it is.

If AI is to become more humanlike in its abilities, the machine-learning and neuroscience communities need to interact closely, something that is happening already. Some of today's greatest exponents of machine learning—such as Geoffrey Hinton, Zoubin Ghahramani, and Demis Hassabis—have backgrounds in cognitive neuroscience, and their success has been at least in part due to attempts to model brainlike behavior in their algorithms. At the same time, neurobiology has also flourished. All sorts of tools have been developed to watch which neurons are firing and genetically manipulate them and see what's happening in real time with inputs. Several countries have launched moon-shot neuroscience initiatives to see if we can crack the workings of the brain. Advances in AI and neuroscience seem to go hand in hand; each field can propel the other.

Many evolutionary scientists, and such philosophers as Daniel Dennett, have pointed out that the human brain is the result of billions of years of evolution.[35] Human intelligence is not the special characteristic we think it is, but just another survival mechanism not unlike our digestive or immune systems, both of which are also amazingly complex. Intelligence evolved because it allowed us to make sense of the world around us, to plan ahead, and thus cope with all sorts of unexpected things in order to survive. However, as Descartes stated, we humans define our very existence by our ability to think. So it is not surprising that, in an anthropomorphic way, our fears about AI reflect this belief that our intelligence is what makes us special.

But if we step back and look at life on Earth, we see that we are far from the most resilient species. If we're going to be taken over at some point, it will be by some of Earth's oldest life-forms, like bacteria, which can live anywhere from Antarctica to deep-sea thermal vents hotter than boiling water, or in acid environments that would melt you and me. So when people ask where we're headed, we need to put the question in a broader context. I don't know what sort of future AI will bring: whether AI will make humans subservient or obsolete or will be a useful and welcome enhancement of our abilities which will enrich our lives. But I am reasonably certain that computers will never be the overlords of bacteria.

---

[35] See, for example, Dennett's *From Bacteria to Bach and Back: The Evolution of Minds* (New York: W. W. Norton, 2017).

Alex "Sandy" Pentland, *an exponent of what he has termed "social physics," is interested in building powerful human-AI ecologies. He is concerned at the same time about the potential dangers of decision-making systems in which the data in effect take over and human creativity is relegated to the background.*

The advent of Big Data, he believes, has given us the opportunity to reinvent our civilization: "We can now begin to actually look at the details of social interaction and how those play out, and we're no longer limited to averages like market indices or election results. This is an astounding change. The ability to see the details of the market, of political revolutions, and to be able to predict and control them is definitely a case of Promethean fire—it could be used for good or for ill. Big Data brings us to interesting times."

At our group meeting in Washington, Connecticut, he confessed that reading Norbert Wiener on the concept of feedback "felt like reading my own thoughts."

"After Wiener, people discovered or focused on the fact that there are genuinely chaotic systems that are just not predictable," he said, "but if you look at human socioeconomic systems, there is a large percentage of variance you can account for and predict. . . . Today there is data from all sorts of digital devices, and from all of our transactions. The fact that everything is datafied means you can measure things in real time in most aspects of human life—and increasingly in every aspect of human life. The fact that we have interesting computers and machine-learning techniques means that you can build predictive models of human systems in ways you could never do before."

THE HUMAN STRATEGY
Alex “Sandy” Pentland

Alex “Sandy” Pentland *is Toshiba Professor and professor of media arts and sciences, MIT; director of the Human Dynamics and Connection Science labs and the Media Lab Entrepreneurship Program, and the author of* Social Physics.

In the last half-century, the idea of AI and intelligent robots has dominated thinking about the relationship between humans and computers. In part, this is because it’s easy to tell the stories about AI and robots, and in part because of early successes (e.g., theorem provers that reproduced most of Whitehead and Russell’s *Principia Mathematica*) and massive military funding. The earlier and broader vision of cybernetics, which considered the artificial as part of larger systems of feedback and mutual influence, faded from public awareness.

However, in the intervening years the cybernetics vision has slowly grown and quietly taken over—to the point where it is “in the air.” State-of-the-art research in most engineering disciplines is now framed as feedback systems that are dynamic and driven by energy flows. Even AI is being recast as human/machine “advisor” systems, and the military is beginning large-scale funding in this area—something that should perhaps worry us more than drones and independent humanoid robots.

But as science and engineering have adopted a more cybernetics-like stance, it has become clear that even the vision of cybernetics is far too small. It was originally centered on the embeddedness of the individual actor but not on the emergent properties of a network of actors. This is unsurprising, because the mathematics of networks did not exist until recently, so a quantitative science of how networks behave was impossible. We now know that study of the individual does not produce understanding of the system except in certain simple cases. Recent progress in this area was foreshadowed by understanding that “chaos,” and later “complexity,” were the typical behavior of systems, but we can now go far beyond these statistical understandings.

We’re beginning to be able to analyze, predict, and even design the emergent behavior of complex heterogeneous networks. The cybernetics view of the connected individual actor can now be expanded to cover complex systems of connected individuals and machines, and the insights we obtain from this broader view are fundamentally different from those obtained from the cybernetics view. Thinking about the network is analogous to thinking about entire ecosystems. How would you guide ecosystems to grow in a good direction? What do you even mean by “a good direction”? Questions like this are beyond the boundary of traditional cybernetic thinking.

Perhaps the most stunning realization is that humans are already beginning to use AI and machine learning to guide entire ecosystems, including ecosystems of people, thus creating human-AI ecologies. Now that everything is becoming “datafied,” we can measure most aspects of human life and, increasingly, aspects of all life. This, together with new, powerful machine-learning techniques, means that we can build models of these ecologies in ways we couldn’t before. Well-known examples are weather- and traffic-prediction models, which are being extended to predict the global climate and plan city growth and renewal. AI-aided engineering of the ecologies is already here.

Development of human-AI ecosystems is perhaps inevitable for a social species such as ourselves. We became social early in our evolution, millions of years ago. We began exchanging information with one another to stay alive, to increase our fitness. We developed writing to share abstract and complex ideas, and most recently we've developed computers to enhance our communication abilities. Now we're developing AI and machine-learning models of ecosystems and sharing the predictions of those models to jointly shape our world through new laws and international agreements.

We live in an unprecedented historic moment, in which the availability of vast amounts of human behavioral data and advances in machine learning enable us to tackle complex social problems through algorithmic decision making. The opportunities for such a human-AI ecology to have positive social impact through fairer and more transparent decisions are obvious. But there are also risks of a "tyranny of algorithms," where unelected data experts are running the world. The choices we make now are perhaps even more momentous than those we faced in the 1950s, when AI and cybernetics were created. The issues look similar, but they're not. We have moved down the road, and now the scope is larger. It's not just AI robots versus individuals. It's AI guiding entire ecologies.

~ ~ ~

How can we make a good human-artificial ecosystem, something that's not a machine society but a cyberculture in which we can all live as humans—a culture with a human feel to it? We don't want to think small—for example, to talk only of robots and self-driving cars. We want this to be a global ecology. Think Skynet-size. But how would you make Skynet something that's about the human fabric?

The first thing to ask is: What's the magic that makes the current AI work? Where is it wrong and where is it right?

The good magic is that it has something called the credit-assignment function. What that lets you do is take "stupid neurons"—little linear functions—and figure out, in a big network, which ones are doing the work and strengthen them. It's a way of taking a random bunch of switches all hooked together in a network and making them smart by giving them feedback about what works and what doesn't. This sounds simple, but there's some complicated math around it. That's the magic that makes current AI work.

The bad part of it is, because those little neurons are stupid, the things they learn don't generalize very well. If an AI sees something it hasn't seen before, or if the world changes a little bit, the AI is likely to make a horrible mistake. It has absolutely no sense of context. In some ways, it's as far from Norbert Wiener's original notion of cybernetics as you can get, because it isn't contextualized; it's a little idiot savant.

But imagine that you took away those limitations: Imagine that instead of using dumb neurons, you used neurons in which real-world knowledge was embedded. Maybe instead of linear neurons, you used neurons that were functions in physics, and then you tried to fit physics data. Or maybe you put in a lot of knowledge about humans and how they interact with one another—the statistics and characteristics of humans.

When you add this background knowledge and surround it with a good credit-assignment function, then you can take observational data and use the credit-assignment function to reinforce the functions that are producing good answers. The result is an AI that works extremely well and can generalize. For instance, in solving physical problems, it often takes only a couple of noisy data points to get something that's a beautiful description of a phenomenon, because you're putting in knowledge about how physics works. That's in huge contrast to normal AI, which requires millions of training examples and is very sensitive to noise. By adding the appropriate background knowledge, you get much more intelligence.

Similar to the physical-systems case, if we make neurons that know a lot about how humans learn from each other, then we can detect human fads and predict human behavior trends in surprisingly accurate and efficient ways. This "social physics" works because human behavior is determined as much by the patterns of our culture as by rational, individual thinking. These patterns can be described mathematically and employed to make accurate predictions.

This idea of a credit-assignment function reinforcing connections between neurons that are doing the best work is the core of current AI. If you make those little neurons smarter, the AI gets smarter. So, what would happen if we replaced the neurons with people? People have lots of capabilities. They know lots of things about the world; they can perceive things in a broadly competent, human way. What would happen if you had a network of people in which you could reinforce the connections that were helping and minimize the connections that weren't?

That begins to sound like a society, or a company. We all live in a human social network. We're reinforced for doing things that seem to help everybody and discouraged from doing things that are not appreciated. Culture is the result of this sort of human AI as applied to human problems; it is the process of building social structures by reinforcing the good connections and penalizing the bad. Once you've realized you can take this general AI framework and create a human AI, the question becomes, What's the right way to do that? Is it a safe idea? Is it completely crazy?

My students and I are looking at how people make decisions, on huge databases of financial decisions, business decisions, and many other sorts of decisions. What we've found is that humans often make decisions in a way that mimics AI credit-assignment algorithms and works to make the community smarter. A particularly interesting feature of this work is that it addresses a classic problem in evolution known as the group-selection problem. The core of this problem is: How can we select for culture in evolution, when it's the individuals that reproduce? What you need is something that selects for the best cultures and the best groups but also selects for the best individuals, because they're the units that transmit the genes.

When you frame the question this way and go through the mathematical literature, you discover that there's one generally best way to do this. It's called "distributed Thompson sampling," a mathematical algorithm used in choosing, out of a set of possible actions with unknown payoffs, the action that maximizes the expected reward in respect to the actions. The key is social sampling, a way of combining evidence, of exploring and exploiting at the same time. It has the unusual property of simultaneously being the best strategy both for the individual and for the group. If you use the group as the basis of selection, and then the group either gets wiped out or reinforced, you're also selecting for successful individuals. If you select for individuals, and each individual does what's good for him or her, then that's automatically the best thing for the group. It's an amazing alignment of interests and utilities, and it provides real insight into the question of how culture fits into natural selection.

Social sampling, very simply, is looking around you at the actions of people who are like you, finding what's popular, and then copying it if it seems like a good idea to you. Idea propagation has this popularity function driving it, but individual adoption also is about figuring out how the idea works for the individual—a reflective attitude. When you combine social sampling and personal judgment, you get superior decision making. That's amazing, because now we have a mathematical recipe for doing with humans what all those AI techniques are doing with dumb computer neurons. We have a way of putting people together to make better decisions, given more and more experience.

So, what happens in the real world? Why don't we do this all the time? Well, people are good at it, but there are ways it can run amok. One of these is through advertising, propaganda, or "fake news." There are many ways to get people to think something is popular when it's not, and this destroys the usefulness of social sampling. The way you can make groups of people smarter, the way you can make human AI, will work only if you can get feedback to them that's truthful. It must be grounded on whether each person's actions worked for them or not.

That's the key to AI mechanisms, too. What they do is analyze whether they performed correctly. If so, plus one; if not, minus one. We need that truthful feedback to make this human mechanism work well, and we need good ways of knowing about what other people are doing so that we can correctly assess popularity and the likelihood of this being a good choice.

The next step is to build this credit-assignment function, this feedback function, for people, so that we can make a good human-artificial ecosystem—a smart organization and a smart culture. In a way, we need to duplicate some of the early insights that resulted in, for instance, the U.S. census—trying to find basic facts that everybody can agree on and understand so that the transmission of knowledge and culture can happen in a way that's truthful and social sampling can function efficiently.

We can address the problem of building an accurate credit-assignment function in many different settings. In companies, for instance, it can be done with digital ID badges that reveal who's connected to whom, so that we can assess the pattern of connections in relation to the company's results on a daily or weekly basis. The credit-assignment function asks whether those connections helped solve problems, or helped invent new solutions, and reinforces the helpful connections. When you can get that feedback quantitatively—which is difficult, because most things aren't measured quantitatively—both the productivity and the innovation rate within the organization can be significantly improved. This is, for instance, the basis of Toyota's "continuous improvement" method.

A next step is to try to do the same thing but at scale, something I refer to as building a trust network for data. It can be thought of as a distributed system like the Internet, but with the ability to quantitatively measure and communicate the qualities of human society, in the same way that the U.S. census does a pretty good job of telling us about population and life expectancy. We are already deploying prototype examples of trust networks at scale in several countries, based on the data and measurement standards laid out in the U.N. Sustainable Development Goals.

On the horizon is a vision of how we can make humanity more intelligent by building a human AI. It's a vision composed of two threads. One is data that we can all trust—data that have been vetted by a broad community, data where the algorithms are known and monitored, much like the census data we all automatically rely on as at least approximately correct. The other is a fair, data-driven assessment of public norms, policy, and government, based on trusted data about current conditions. This second thread depends on availability of trusted data and so is just beginning to be developed. Trusted data and data-driven assessment of norms, policy, and government together create a credit-assignment function that improves societies' overall fitness and intelligence.

It is precisely at the point of creating greater societal intelligence where fake news, propaganda, and advertising all get in the way. Fortunately, trust networks give us a path forward to building a society more resistant to echo-chamber problems, these fads, these exercises in madness. We have begun to develop a new way of establishing social measurements, in aid of curing some of the ills we see in society today. We're using open data from all sources, encouraging a fair representation of the things people are choosing, in a curated mathematical framework that can stamp out the echoes and the attempts to manipulate us.

On Polarization and Inequality

Extreme polarization and segregation by income are almost everywhere in the world today and threaten to tear governments and civil society apart. Increasingly, the media are becoming adrenaline pushers driven by advertising clicks and failing to deliver balanced facts and reasoned discourse—and the degradation of media is causing people to lose their bearings. They don't know what to believe, and thus they can easily be manipulated. There is a real need to ground our various cultures in trustworthy, data-driven standards that we all agree on, and to be able to know what behaviors and policies work and which don't.

In converting to a digital society, we've lost touch with traditional notions of truth and justice. Justice used to be mostly informal and normative. We've now formalized it. At the same time, we've put it out of reach for most people. Our legal systems are failing us in a way they didn't before, precisely because they're now more formal, more digital, less embedded in society.

Ideas about justice are very different around the world. One of the core differentiators is this: Do you or your parents remember when the bad guys came with guns and took everything? If you do, your attitude about justice is different from that of the average reader of this essay. Do you come from the upper classes? Or were you somebody who saw the sewers from the inside? Your view of justice depends on your history.

A common test I have for U.S. citizens is this: Do you know anybody who owns a pickup truck? It's the number-one-selling vehicle in the United States, and if you don't know people like that, you're out of touch with more than 50 percent of Americans. Physical segregation drives conceptual segregation. Most of America thinks of justice and access and fairness in terms very different from those of the typical, say, Manhattanite.

If you look at patterns of mobility—where people go—in a typical city, you find that the people in the top quintile (white-collar working families) and bottom quintile (people who are sometimes on unemployment or welfare) almost never talk to each other. They don't go to the same places; they don't talk about the same things. They all live in the same city, nominally, but it's as if it were two completely different cities—and this is perhaps the most important cause of today's plague of polarization.

On Extreme Wealth

Some two hundred of the world's wealthiest people have pledged to give away more than 50 percent of their wealth either during their lifetimes or in their wills, creating a plurality of voices in the foundation space.[36] Bill Gates is probably the most familiar example. He's decided that if the government won't do it, he'll do it. You want mosquito nets? He'll do it. You want antivirals? He'll do it. We're getting different stakeholders to take action in the form of foundations dedicated to public good, and they have different versions of what they consider the public good. This diversity of goals has created a lot of what's wonderful about the world today. Actions from outside government by organizations like the Ford Foundation and the Sloan Foundation, who bet on things that nobody else would bet on, have changed the world for the better.

Sure, these billionaires are human, with human foibles, and all is not necessarily as it should be. On the other hand, the same situation obtained when the railways were first built. Some people made huge fortunes. A lot of people went bust. We, the average people, got railways out of it. That's good. Same thing with electric power; same thing with many new technologies. There's a churning process that throws somebody up and later casts them or their heirs down. Bubbles of extreme wealth were a feature of the late 1800s and early 1900s when steam engines and railways and electric lights were invented. The fortunes they created were all gone within two or three generations.

If the U.S. were like Europe, I would worry. What you find in Europe is that the same families have held on to wealth for hundreds of years, so they're entrenched not just in terms of wealth but of the political system and in other ways. But so far, the U.S. has avoided this kind of hereditary class system. Extreme wealth hasn't stuck, which is good. It shouldn't stick. If you win the lottery, you get your billion dollars, but your grandkids ought to work for a living.

On AI and Society

People are scared about AI. Perhaps they should be. But they need to realize that AI feeds on data. Without data, AI is nothing. You don't have to watch the AI; instead you should watch what it eats and what it does. The trust-network framework we've set up, with the help of nations in the E.U. and elsewhere, is one where we can have our algorithms, we can have our AI, but we get to see what went in and what went out, so that we can ask, Is this a discriminatory decision? Is this the sort of thing that we want as humans? Or is this something that's a little weird?

The most revealing analogy is that regulators, bureaucracies, and parts of the government are very much like AIs: They take in the rules that we call law and regulation, and they add government data, and they make decisions that affect our lives. The part that's bad about the current system is that we have very little oversight of these departments, regulators, and bureaucracies. The only control we have is the vote—the opportunity to elect somebody different. We need to make oversight of bureaucracies a lot more fine-grained. We need to record the data that went into every single decision and have the results analyzed by the various stakeholders—rather like elected legislatures were originally intended to do.

---

[36] https://givingpledge.org/About.aspx.

If we have the data that go into and out of each decision, we can easily ask, Is this a fair algorithm? Is this AI doing things that we as humans believe are ethical? This human-in-the-loop approach is called "open algorithms;" you get to see what the AIs take as input and what they decide using that input. If you see those two things, you'll know whether they're doing the right thing or the wrong thing. It turns out that's not hard to do. If you control the data, then you control the AI.

One thing people often fail to mention is that all the worries about AI are the same as the worries about today's government. For most parts of the government—the justice system, et cetera—there's no reliable data about what they're doing and in what situation. How can you know whether the courts are fair or not if you don't know the inputs and the outputs? The same problem arises with AI systems and is addressable in the same way. We need trusted data to hold current government to account in terms of what they take in and what they put out, and AI should be no different.

Next-Generation AI

Current AI machine-learning algorithms are, at their core, dead simple stupid. They work, but they work by brute force, so they need hundreds of millions of samples. They work because you can approximate anything with lots of little simple pieces. That's a key insight of current AI research—that if you use reinforcement learning for credit-assignment feedback, you can get those little pieces to approximate whatever arbitrary function you want.

But using the wrong functions to make decisions means the AI's ability to make good decisions won't generalize. If we give the AI new, different inputs, it may make completely unreasonable decisions. Or if the situation changes, then you need to retrain it. There are amusing techniques to find the "null space" in these AI systems. These are inputs that the AI thinks are valid examples of what it was trained to recognize (e.g., faces, cats, etc.), but to a human they're crazy examples.

Current AI is doing descriptive statistics in a way that's not science and would be almost impossible to make into science. To build robust systems, we need to know the science behind data. The systems I view as next-generation AIs result from this science-based approach: If you're going to create an AI to deal with something physical, then you should build the laws of physics into it as your descriptive functions, in place of those stupid little neurons. For instance, we know that physics uses functions like polynomials, sine waves, and exponentials, so those should be your basis functions and not little linear neurons. By using those more appropriate basis functions, you need a lot less data, you can deal with a lot more noise, and you get much better results.

As in the physics example, if we want to build an AI to work with human behavior, then we need to build the statistical properties of human networks into machine-learning algorithms. When you replace the stupid neurons with ones that capture the basics of human behavior, then you can identify trends with very little data, and you can deal with huge levels of noise.

The fact that humans have a "commonsense" understanding that they bring to most problems suggests what I call the human strategy: Human society is a network just like the neural nets trained for deep learning, but the "neurons" in human society are a lot smarter. You and I have surprisingly general descriptive powers that we use for understanding a wide range of situations, and we can recognize which connections should be reinforced. That means we can shape our social networks to work much better and potentially beat all that machine-based AI at its own game.

"URGENT!" URGENT!" the cc'd copy of an email screamed, one of a dozen emails that greeted me as I turned on my phone at the baggage carousel at Malpensa Airport after the long flight from JFK. "The great American visionary thinker John Brockman arrives this morning at Grand Hotel Milan. You MUST, repeat MUST pay him a visit." It was signed HUO.

The prior evening, waiting in the lounge at JFK, I had had the bright idea to write my friend and longtime collaborator, the London-based, peripatetic art curator **Hans Ulrich Obrist** *(known to all as HUO), and ask if there was anyone in Milan I should know.*

Once I was settled at the hotel, the phone began ringing and a procession of leading Italian artists, designers, and architects called to request a meeting, including Enzo Mari, the modernist artist and furniture designer; Alberto Garutti, whose aesthetic strategies have inspired a dialogue between contemporary art, spectator, and public space; and fashion designer Miuccia Prada, who "requests your presence for tea this afternoon at Prada headquarters." And thus, thanks to HUO, did the jet-lagged "great American visionary thinker" stumble and mumble his way through his first day in Milan, November 2011.

HUO is sui generis*: He lives a twenty-four-hour day, sleeping (I guess) whenever, and employing full-time assistants who work eight-hour shifts and are available to him 24/7. Over a recent two-year period, he visited art venues in either China or India for forty weekends each year—departing London Thursday evening, back at his desk on Monday. Last year, once again,* ArtReview *ranked him #1 on their annual "Power 100" list.*

Recently we collaborated on a panel during the "GUEST, GHOST, HOST: MACHINE!" Serpentine event that took place at London's new City Hall. We were joined by Venki Ramakrishnan, Jaan Tallinn, and Andrew Blake, research director of The Alan Turing Institute. The event was consistent with HUO's mission of bringing together art and science: "The curator is no longer understood simply as the person who fills a space with objects," he says, "but also as the person who brings different cultural spheres into contact, invents new display features, and makes junctions that allow unexpected encounters and results."

MAKING THE INVISIBLE VISIBLE: ART MEETS AI

Hans Ulrich Obrist

***Hans Ulrich Obrist** is artistic director of the Serpentine Gallery, London, and the author of* Ways of Curating *and* Lives of the Artists, Lives of the Architects.

In the Introduction to the second edition of his book *Understanding Media*, Marshall McLuhan noted the ability of art to "anticipate future social and technological developments." Art is "an early alarm system," pointing us to new developments in times ahead and allowing us "to prepare to cope with them. . . . Art as a radar environment takes on the function of indispensable perceptual training. . . ."

In 1964, when McLuhan's book was first published, the artist Nam June Paik was just building his *Robot K-456* to experiment with the technologies that subsequently would start to influence society. He had worked with television earlier, challenging its usual passive consumption by the viewer, and later made art with global live-satellite broadcasts, using the new media less for entertainment than to point us to their poetic and intercultural capacities (which are still mostly unused today). The Paiks of our time, of course, are now working with the Internet, digital images, and artificial intelligence. Their works and thoughts, again, are an early alarm system for the developments ahead of us.

As a curator, my daily work is to bring together different works of art and connect different cultures. Since the early 1990s, I have also been organizing conversations and meetings with practitioners from different disciplines, in order to go beyond the general reluctance to pool knowledge. Since I was interested in hearing what artists have to say about artificial intelligence, I recently organized several conversations between artists and engineers.

The reason to look closely at AI is that two of the most important questions of today are "How capable will AI become?" and "What dangers may arise from it?" Its early applications already influence our everyday lives in ways that are more or less recognizable. There is an increasing impact on many aspects of our society, but whether this might be, in general, beneficial or malign is still uncertain.

Many contemporary artists are following these developments closely. They are articulating various doubts about the promises of AI and reminding us not to associate the term "artificial intelligence" solely with positive outcomes. To the current discussions of AI, the artists contribute their specific perspectives and notably their focus on questions of image making, creativity, and the use of programming as artistic tools.

The deep connections between science and art had already been noted by the late Heinz von Foerster, one of the architects of cybernetics, who worked with Norbert Wiener from the mid-1940s and in the 1960s founded the field of second-order cybernetics, in which the observer is understood as part of the system itself and not an external entity. I knew von Foerster well, and in one of our many conversations, he offered his views on the relation between art and science:

> I've always perceived art and science as complementary fields. One shouldn't forget that a scientist is in some respects also an artist. He invents a new technique and he describes it. He uses language like a poet, or the author of a detective novel, and describes his findings. In my view, a scientist must work in

> an artistic way if he wants to communicate his research. He obviously wants to communicate and talk to others. A scientist invents new objects, and the question is how to describe them. In all of these aspects, science is not very different from art.

When I asked him how he defined cybernetics, von Foerster answered:

> The substance of what we have learned from cybernetics is to think in circles: *A* leads *B, B* to *C,* but *C* can return to *A*. Such kinds of arguments are not linear but circular. The significant contribution of cybernetics to our thinking is to accept circular arguments. This means that we have to look at circular processes and understand under which circumstances an equilibrium, and thus a stable structure, emerges.

Today, where AI algorithms are applied in daily tasks, one can ask how the human factor is included in these kinds of processes and what role creativity and art could play in relation to them. There are thus different levels to think about when exploring the relation between AI and art.

So, what do contemporary artists have to say about artificial intelligence?

Artificial Stupidity

Hito Steyerl, an artist who works with documentary and experimental film, considers two key aspects that we should keep in mind when reflecting on the implications of AI for society. First, the expectations for so-called artificial intelligence, she says, are often overrated, and the noun "intelligence" is misleading; to counter that, she uses the term "artificial stupidity." Second, she points out that programmers are now making invisible software algorithms visible through images, but to understand and interpret these images better, we should apply the expertise of artists.

Steyerl has worked with computer technology for many years, and her recent artworks have explored surveillance techniques, robots, and such computer games as in *How Not to Be Seen* (2013), on digital-image technologies, or *HellYeahWeFuckDie* (2017), about the training of robots in the still-difficult task of keeping balance. But to explain her notion of artificial stupidity, Steyerl refers to a more general phenomenon, like the now widespread use of Twitter bots, noting in our conversation:

> It was and still is a very popular tool in elections to deploy Twitter armies to sway public opinion and deflect popular hashtags and so on. This is an artificial intelligence of a very, very low grade. It's two or maybe three lines of script. It's nothing very sophisticated at all. Yet the social implications of this kind of artificial stupidity, as I call it, are already monumental in global politics.

As has been widely noted, this kind of technology was seen in the many automated Twitter posts before the 2016 U.S. presidential election and also shortly before the Brexit vote. If even low-grade AI technology like these bots are already influencing our politics, this raises another urgent question: "How powerful will far more advanced techniques be in the future?"

Visible / Invisible

The artist Paul Klee often talked about art as "making the invisible visible." In computer technology, most algorithms work invisibly, in the background; they remain inaccessible in the systems we use daily. But lately there has been an interesting comeback of visuality in machine learning. The ways that the deep-learning algorithms of AI are processing data have been made visible through applications like Google's DeepDream, in which the process of computerized pattern-recognition is visualized in real time. The application shows how the algorithm tries to match animal forms with any given input. There are many other AI visualization programs that, in their way, also "make the invisible visible." The difficulty in the general public perception of such images is, in Steyerl's view, that these visual patterns are viewed uncritically as realistic and objective representations of the machine process. She says of the aesthetics of such visualizations:

> For me, this proves that science has become a subgenre of art history. . . . We now have lots of abstract computer patterns that might look like a Paul Klee painting, or a Mark Rothko, or all sorts of other abstractions that we know from art history. The only difference, I think, is that in current scientific thought they're perceived as representations of reality, almost like documentary images, whereas in art history there's a very nuanced understanding of different kinds of abstraction.

What she seeks is a more profound understanding of computer-generated images and the different aesthetic forms they use. They are obviously not generated with the explicit goal of following a certain aesthetic tradition. The computer engineer Mike Tyka, in a conversation with Steyerl, explained the functions of these images:

> Deep-learning systems, especially the visual ones, are really inspired by the need to know what's going on in the black box. Their goal is to project these processes back into the real world.

Nevertheless, these images have aesthetic implications and values which have to be taken into account. One could say that while the programmers use these images to help us better understand the programs' algorithms, we need the knowledge of artists to better understand the aesthetic forms of AI. As Steyerl has pointed out, such visualizations are generally understood as "true" representations of processes, but we should pay attention to their respective aesthetics, and their implications, which have to be viewed in a critical and analytical way.

In 2017, the artist Trevor Paglen created a project to make these invisible AI algorithms visible. In *Sight Machine*, he filmed a live performance of the Kronos Quartet and processed the resulting images with various computer software programs used for face detection, object identification, and even for missile guidance. He projected the outcome of these algorithms, in real time, back to screens above the stage. By demonstrating how the various different programs interpreted the musicians' performance, Paglen showed that AI algorithms are always determined by sets of values and interests which they then manifest and reiterate, and thus must be critically questioned. The significant contrast between algorithms and music also raises the issue of relationships between technical and human perception.

Computers, as a Tool for Creativity, Can't Replace the Artist.

Rachel Rose, a video artist who thinks about the questions posed by AI, employs computer technology in the creation of her works. Her films give the viewer an experience of materiality through the moving image. She uses collaging and layering of the material to manipulate sound and image, and the editing process is perhaps the most important aspect of her work.

She also talks about the importance of decision making in her work. For her, the artistic process does not follow a rational pattern. In a converation we had, together with the engineer Kenric McDowell, at the Google Cultural Institute, she explained this by citing a story from theater director Peter Brook's 1968 book *The Empty Space*. When Brook designed the set for his production of *The Tempest* in the late 1960s, he started by making a Japanese garden, but then the design evolved, becoming a white box, a black box, a realistic set, and so on. And in the end, he returned to his original idea. Brook writes that he was shocked at having spent a month on his labors, only to end at the beginning. But this shows that the creative artistic process is a succession whose every step builds on the next and which eventually comes to an unpredictable conclusion. The process is not a logical or rational succession but has mostly to do with the artist's feelings in reaction to the preceding result. Rose said, of her own artistic decision making:

> It, to me, is distinctively different from machine learning, because at each decision there's this core feeling that comes from a human being, which has to do with empathy, which has to do with communication, which has to do with questions about our own mortality that only a human could ask.

This point underlines the fundamental difference between any human artistic production and so-called computer creativity. Rose sees AI more as a possible way to create better tools for humans:

> A place I can imagine machine learning working for an artist would be not in developing an independent subjectivity, like writing a poem or making an image, but actually in filling in gaps that are to do with labor, like the way that Photoshop works with different tools that you can use.

And though such tools may not seem spectacular, she says, "they might have a larger influence on art," because they provide artists with further possibilities in their creative work.

McDowell added that he, too, believes there are false expectations around AI. "I've observed," he said, "that there's a sort of magical quality to the idea of a computer that does all the things that we do." He continued: "There's almost this kind of demonic mirror that we look into, and we want it to write a novel, we want it to make a film—we want to give that away somehow." He is instead working on projects wherein humans collaborate with the machine. One of the current aims of AI research is to find new means of interaction between humans and software. And art, one could say, needs to play a key role in that enterprise, since it focuses on our subjectivity and on essential human aspects like empathy and mortality.

Cybernetics / Art

Suzanne Treister is an artist whose work from 2009 to 2011 serves as an example of what is happening at the intersection of our current technologies, the arts, and cybernetics. Treister has been a pioneer in digital art since the 1990s, inventing, for example, imaginary video games and painting screen shots from them. In her project *Hexen 2.0* she looked back at the famous Macy conferences on cybernetics that between 1946 and 1953 were organized in New York by engineers and social scientists to unite the sciences and to develop a universal theory of the workings of the mind.

In her project, she created thirty photo-text works about the conference attendees (which included Wiener and von Foerster), she invented tarot cards, and she made a video based on a photomontage of a "cybernetic séance." In the "séance," the conference participants are seen sitting at a round table, as in spiritualist séances, while certain of their statements on cybernetics are heard in an audio-collage—rational knowledge and superstition combined. She also noted that some of the participating scientists worked for the military; thus the application of cybernetics could be seen in an ambivalent way, even back then, as a tussle between pure knowledge and its use in state control.

If one looks at Treister's work about the Macy conference participants, one sees that no visual artist was included. A dialogue between artists and scientists would be fruitful in future discussions, and it is a bit astonishing that this wasn't realized at the time, given von Foerster's keen interest in art. He recounted in one of our conversations how his relation to the field dated back to his childhood:

> I grew up as a child in an artistic family. We often had visits from poets, philosophers, painters, and sculptors. Art was a part of my life. Later, I got into physics, as I was talented in this subject. But I always remained conscious of the importance of art for science. There wasn't a great difference for me. For me, both aspects of life have always been very much alike—and accessible, too. We should see them as one. An artist also has to reflect on his work. He has to think about his grammar and his language. A painter must know how to handle his colors. Just think of how intensively oil colors were researched during the Renaissance. They wanted to know how a certain pigment could be mixed with others to get a certain tone of red or blue. Chemists and painters collaborated very closely. I think the artificial division between science and art is wrong.

Though for von Foerster the relation between the art and science was always clear, for our own time this connection remains to be made. There are many reasons to multiply the links. The critical thinking of artists would be beneficial in respect to the dangers of AI, since they draw our attention to questions they consider essential from their perspective. With the advent of machine learning, new tools are available to artists for their work. And as the algorithms of AI are made visible through artificial images in new ways, artists' critical visual knowledge and expertise will be harnessed. Many of the key questions of AI are philosophical in nature and can be answered only from a holistic point of view. The way they play out among adventurous artists will be worth following.

Simulating Worlds

For the most part, the works of contemporary artists have been embodied ruminations on

AI's impact on existential questions of the self and our future interaction with nonhuman entities. Few, though, have taken the technologies and innovations of AI as the underlying materials of their work and sculpted them to their own vision. An exception is the artist Ian Cheng, who has gone as far as to construct entire worlds of artificial beings with varying degrees of sentience and intelligence. He refers to these worlds as Live Simulations. His *Emissaries* trilogy (2015-2017) is set in a fictional postapocalyptic world of flora and fauna, in which AI-driven animals and creatures explore the landscape and interact with each other. Cheng uses advanced graphics but has them programmed with a lot of glitches and imperfections, which imparts a futuristic and anachronistic atmosphere at the same time. Through his trilogy, which charts a history of consciousness, he asks the question "What is a simulation?"

While the majority of artistic works that utilize recent developments in AI specifically draw from the field of machine learning, Cheng's Live Simulations take a separate route. The protagonists and plot lines that are interlaced in each episodic simulation of *Emissaries* use the complex logic systems and rules of AI. What is profound about his continually evolving scenes is that complexity arises not through the desire/actions of any single actor or artificial godhead but instead through their constellation, collision, and constant evolution in symbiosis with one another. This gives rise to unexpected outcomes and unending, unknowable situations—you can never experience the exact same moment in successive viewings of his work.

Cheng had a discussion at the Serpentine Marathon "GUEST, GHOST, HOST: MACHINE!" with the programmer Richard Evans, who recently designed Versu, an AI-based platform for interactive storytelling games. Evans' work emphasizes the social interaction of the games' characters, who react in a spectrum of possible behaviors to the choices made by the human players. In their conversation, Evans said that a starting point for the project was that most earlier simulation video games, such as *The Sims*, did not sufficiently take into account the importance of social practices. Simulated protagonists in games would often act in ways that did not correspond well with real human behavior. Knowledge of social practices limits the possibilities of action but is necessary to understand the meaning of our actions—which is what interests Cheng for his own simulations. The more parameters of actions in certain circumstances are determined in a computer simulation, the more interesting it is for Cheng to experiment with individual and specific changes. He told Evans, "I gather that if we had AI with more ability to respond to social contexts, tweaking one thing, you would get something quite artistic and beautiful."

Cheng also sees the work of programmers and AI simulations as creating new and sophisticated tools for experimenting with the parameters of our daily social practices. In this way, the involvement of artists in AI will lead to new kinds of open experiments in Art. Such possibilities are—like increased AI capabilities in general—still in the future. Recognizing that this is an experimental technology in its infancy, very far from apocalyptic visions of a superintelligent AI takeover, Cheng fills his simulations with prosaic avatars such as strange microbial globules, dogs, and the undead.

Discussions like these, between artists and engineers, of course are not totally new. In the 1960s, the engineer Billy Klüver brought artists together with engineers in a series of events, and in 1967 he founded the Experiments in Art and Technology program with Robert Rauschenberg and others. In London, at around the same time, Barbara

Stevini and John Latham, of the Artist Placement Group, took things a step further by asserting that there should be artists in residence in every company and every government. Today, these inspiring historical models can be applied to the field of AI. As AI comes to inhabit more and more of our everyday lives, the creation of a space that is nondeterministic and non-utilitarian in its plurality of perspectives and diversity of understandings will undoubtedly be essential.

Alison Gopnik *is an international leader in the field of children's learning and development and was one of the founders of the field of "theory of mind." She has spoken of the child brain as a "powerful learning computer," perhaps from personal experience. Her own Philadelphia childhood was an exercise in intellectual development. "Other families took their kids to see* The Sound of Music *or* Carousel*; we saw Racine's* Phaedra *and Samuel Beckett's* Endgame," *she has recalled. "Our family read Henry Fielding's 18th-century novel* Joseph Andrews *out loud to each other around the fire on camping trips."*

Lately she has invoked Bayesian models of machine learning to explain the remarkable ability of preschoolers to draw conclusions about the world around them without benefit of enormous data sets. "I think babies and children are actually more conscious than we are as adults," she has said. "They're very good at taking in lots of information from lots of different sources at once." She has referred to babies and young children as "the research and development division of the human species." Not that she treats them coldly, as if they were mere laboratory animals. They appear to revel in her company, and in the blinking, thrumming toys in her Berkeley lab. For years after her own children had outgrown it, she kept a playpen in her office.

Her investigations into just how we learn, and the parallels to the deep-learning methods of AI, continues. "It turns out to be much easier to simulate the reasoning of a highly trained adult expert than to mimic the ordinary learning of every baby," she says. "Computation is still the best—indeed, the only—scientific explanation we have of how a physical object like a brain can act intelligently. But, at least for now, we have almost no idea at all how the sort of creativity we see in children is possible."

AIs VERSUS FOUR-YEAR-OLDS

Alison Gopnik

***Alison Gopnik** is a developmental psychologist at UC Berkeley; her books include* The Philosophical Baby *and, most recently,* The Gardener and the Carpenter: What the New Science of Child Development Tells Us About the Relationship Between Parents and Children.

Everyone's heard about the new advances in artificial intelligence, and especially machine learning. You've also heard utopian or apocalyptic predictions about what those advances mean. They have been taken to presage either immortality or the end of the world, and a lot has been written about both those possibilities. But the most sophisticated AIs are still far from being able to solve problems that human four-year-olds accomplish with ease. In spite of the impressive name, artificial intelligence largely consists of techniques to detect statistical patterns in large data sets. There is much more to human learning.

How can we possibly know so much about the world around us? We learn an enormous amount even when we are small children; four-year-olds already know about plants and animals and machines; desires, beliefs, and emotions; even dinosaurs and spaceships.

Science has extended our knowledge about the world to the unimaginably large and the infinitesimally small, to the edge of the universe and the beginning of time. And we use that knowledge to make new classifications and predictions, imagine new possibilities, and make new things happen in the world. But all that reaches any of us from the world is a stream of photons hitting our retinas and disturbances of air at our eardrums. How do we learn so much about the world when the evidence we have is so limited? And how do we do all this with the few pounds of grey goo that sits behind our eyes?

The best answer so far is that our brains perform computations on the concrete, particular, messy data arriving at our senses, and those computations yield accurate representations of the world. The representations seem to be structured, abstract, and hierarchical; they include the perception of three-dimensional objects, the grammars that underlie language, and mental capacities like "theory of mind," which lets us understand what other people think. Those representations allow us to make a wide range of new predictions and imagine many new possibilities in a distinctively creative human way.

This kind of learning isn't the only kind of intelligence, but it's a particularly important one for human beings. And it's the kind of intelligence that is a specialty of young children. Although children are dramatically bad at planning and decision making, they are the best learners in the universe. Much of the process of turning data into theories happens before we are five.

Since Aristotle and Plato, there have been two basic ways of addressing the problem of how we know what we know, and they are still the main approaches in machine learning. Aristotle approached the problem from the bottom up: Start with senses—the stream of photons and air vibrations (or the pixels or sound samples of a digital image or recording)—and see if you can extract patterns from them. This approach was carried further by such classic associationists as philosophers David Hume and J. S. Mill and later by behavioral psychologists, like Pavlov and B. F. Skinner. On this view, the abstractness and hierarchical structure of representations is something of an illusion, or at least an epiphenomenon. All the work can be done by association and pattern detection—especially if there are enough data.

Over time, there has been a seesaw between this bottom-up approach to the mystery of learning and Plato's alternative, top-down one. Maybe we get abstract knowledge from concrete data because we already know a lot, and especially because we already have an array of basic abstract concepts, thanks to evolution. Like scientists, we can use those concepts to formulate hypotheses about the world. Then, instead of trying to extract patterns from the raw data, we can make predictions about what the data should look like if those hypotheses are right. Along with Plato, such "rationalist" philosophers and psychologists as Descartes and Noam Chomsky took this approach.

Here's an everyday example that illustrates the difference between the two methods: solving the spam plague. The data consist of a long unsorted list of messages in your in-box. The reality is that some of these messages are genuine and some are spam. How can you use the data to discriminate between them?

Consider the bottom-up technique first. You notice that the spam messages tend to have particular features: a long list of addressees, origins in Nigeria, references to million-dollar prizes or Viagra. The trouble is that perfectly useful messages might have these features, too. If you looked at enough examples of spam and non-spam emails, you might see not only that spam emails tend to have those features but that the features tend to go together in particular ways (Nigeria plus a million dollars spells trouble). In fact, there might be some subtle higher-level correlations that discriminate the spam messages from the useful ones—a particular pattern of misspellings and IP addresses, say. If you detect those patterns, you can filter out the spam.

The bottom-up machine-learning techniques do just this. The learner gets millions of examples, each with some set of features and each labeled as spam (or some other category) or not. The computer can extract the pattern of features that distinguishes the two, even if it's quite subtle.

How about the top-down approach? I get an email from the editor of the *Journal of Clinical Biology*. It refers to one of my papers and says that they would like to publish an article by me. No Nigeria, no Viagra, no million dollars; the email doesn't have any of the features of spam. But by using what I already know, and thinking in an abstract way about the process that produces spam, I can figure out that this email is suspicious.

(1) I know that spammers try to extract money from people by appealing to human greed.

(2) I also know that legitimate "open access" journals have started covering their costs by charging authors instead of subscribers, and that I don't practice anything like clinical biology.

Put all that together and I can produce a good new hypothesis about where that email came from. It's designed to sucker academics into paying to "publish" an article in a fake journal. The email was a result of the same dubious process as the other spam emails, even though it looked nothing like them. I can draw this conclusion from just one example, and I can go on to test my hypothesis further, beyond anything in the email itself, by googling the "editor."

In computer terms, I started out with a "generative model" that includes abstract concepts like greed and deception and describes the process that produces email scams. That lets me recognize the classic Nigerian email spam, but it also lets me imagine many different kinds of possible spam. When I get the journal email, I can work backward: "This seems like just the kind of mail that would come out of a spam-generating process."

The new excitement about AI comes because AI researchers have recently produced powerful and effective versions of both these learning methods. But there is nothing profoundly new about the methods themselves.

Bottom-up Deep Learning

In the 1980s, computer scientists devised an ingenious way to get computers to detect patterns in data: connectionist, or neural-network, architecture (the "neural" part was, and still is, metaphorical). The approach fell into the doldrums in the '90s but has recently been revived with powerful "deep-learning" methods like Google's DeepMind.

For example, you can give a deep-learning program a bunch of Internet images labeled "cat," others labeled "house," and so on. The program can detect the patterns differentiating the two sets of images and use that information to label new images correctly. Some kinds of machine learning, called unsupervised learning, can detect patterns in data with no labels at all; they simply look for clusters of features—what scientists call a factor analysis. In the deep-learning machines, these processes are repeated at different levels. Some programs can even discover relevant features from the raw data of pixels or sounds; the computer might begin by detecting the patterns in the raw image that correspond to edges and lines and then find the patterns in those patterns that correspond to faces, and so on.

Another bottom-up technique with a long history is reinforcement learning. In the 1950s, B. F. Skinner, building on the work of John Watson, famously programmed pigeons to perform elaborate actions—even guiding air-launched missiles to their targets (a disturbing echo of recent AI) by giving them a particular schedule of rewards and punishments. The essential idea was that actions that were rewarded would be repeated and those that were punished would not, until the desired behavior was achieved. Even in Skinner's day, this simple process, repeated over and over, could lead to complex behavior. Computers are designed to perform simple operations over and over on a scale that dwarfs human imagination, and computational systems can learn remarkably complex skills in this way.

For example, researchers at Google's DeepMind used a combination of deep learning and reinforcement learning to teach a computer to play Atari video games. The computer knew nothing about how the games worked. It began by acting randomly and got information only about what the screen looked like at each moment and how well it had scored. Deep learning helped interpret the features on the screen, and reinforcement learning rewarded the system for higher scores. The computer got very good at playing several of the games, but it also completely bombed on others just as easy for humans to master.

A similar combination of deep learning and reinforcement learning has enabled the success of DeepMind's AlphaZero, a program that managed to beat human players at both chess and Go, equipped only with a basic knowledge of the rules of the game and some planning capacities. AlphaZero has another interesting feature: It works by playing hundreds of millions of games against itself. As it does so, it prunes mistakes that led to losses, and it repeats and elaborates on strategies that led to wins. Such systems, and others involving techniques called generative adversarial networks, generate data as well as observing data.

When you have the computational power to apply those techniques to very large data sets or millions of email messages, Instagram images, or voice recordings, you can solve problems that seemed very difficult before. That's the source of much of the excitement in computer science. But it's worth remembering that those problems—like recognizing that an image is a cat or a spoken word is "Siri"—are trivial for a human toddler. One of the most interesting discoveries of computer science is that problems that are easy for us (like identifying cats) are hard for computers—much harder than playing chess or Go. Computers need millions of examples to categorize objects that we can categorize with just a few. These bottom-up systems can generalize to new examples; they can label a new image as a "cat" fairly accurately, over all. But they do so in ways quite different from how humans generalize. Some images almost identical to a cat image won't be identified by us as cats at all. Others that look like a random blur will be.

Top-down Bayesian Models

The top-down approach played a big role in early AI, and in the 2000s it, too, experienced a revival, in the form of probabilistic, or Bayesian, generative models.

The early attempts to use this approach faced two kinds of problems. First, most patterns of evidence might in principle be explained by many different hypotheses: It's possible that my journal email message is genuine, it just doesn't seem likely. Second, where do the concepts that the generative models use come from in the first place? Plato and Chomsky said you were born with them. But how can we explain how we learn the latest concepts of science? Or how even young children understand about dinosaurs and rocket ships?

Bayesian models combine generative models and hypothesis testing with probability theory, and they address these two problems. A Bayesian model lets you calculate just how likely it is that a particular hypothesis is true, given the data. And by making small but systematic tweaks to the models we already have, and testing them against the data, we can sometimes make new concepts and models from old ones. But these advantages are offset by other problems. The Bayesian techniques can help you choose which of two hypotheses is more likely, but there are almost always an enormous number of possible hypotheses, and no system can efficiently consider them all. How do you decide which hypotheses are worth testing in the first place?

Brenden Lake at NYU and colleagues have used these kinds of top-down methods to solve another problem that's easy for people but extremely difficult for computers: recognizing unfamiliar handwritten characters. Look at a character on a Japanese scroll. Even if you've never seen it before, you can probably tell if it's similar to or different from a character on another Japanese scroll. You can probably draw it and even design a fake Japanese character based on the one you see—one that will look quite different from a Korean or Russian character. [37]

---

[37] Brenden M. Lake, Ruslan Salakhutdinov & Joshua B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science,* 350:6266, pp. 1332-38 (2015).

The bottom-up method for recognizing handwritten characters is to give the computer thousands of examples of each one and let it pull out the salient features. Instead, Lake *et al.* gave the program a general model of how you draw a character: A stroke goes either right or left; after you finish one, you start another; and so on. When the program saw a particular character, it could infer the sequence of strokes that were most likely to have led to it—just as I inferred that the spam process led to my dubious email. Then it could judge whether a new character was likely to result from that sequence or from a different one, and it could produce a similar set of strokes itself. The program worked much better than a deep-learning program applied to exactly the same data, and it closely mirrored the performance of human beings.

These two approaches to machine learning have complementary strengths and weaknesses. In the bottom-up approach, the program doesn't need much knowledge to begin with, but it needs a great deal of data, and it can generalize only in a limited way. In the top-down approach, the program can learn from just a few examples and make much broader and more varied generalizations, but you need to build much more into it to begin with. A number of investigators are currently trying to combine the two approaches, using deep learning to implement Bayesian inference.

The recent success of AI is partly the result of extensions of those old ideas. But it has more to do with the fact that, thanks to the Internet, we have much more data, and thanks to Moore's Law we have much more computational power to apply to that data. Moreover, an unappreciated fact is that the data we do have has already been sorted and processed by human beings. The cat pictures posted to the Web are canonical cat pictures—pictures that humans have already chosen as "good" pictures. Google Translate works because it takes advantage of millions of human translations and generalizes them to a new piece of text, rather than genuinely understanding the sentences themselves.

But the truly remarkable thing about human children is that they somehow combine the best features of each approach and then go way beyond them. Over the past fifteen years, developmentalists have been exploring the way children learn structure from data. Four-year-olds can learn by taking just one or two examples of data, as a top-down system does, and generalizing to very different concepts. But they can also learn new concepts and models from the data itself, as a bottom-up system does.

For example, in our lab we give young children a "blicket detector"—a new machine to figure out, one they've never seen before. It's a box that lights up and plays music when you put certain objects on it but not others. We give children just one or two examples of how the machine works, showing them that, say, two red blocks make it go, while a green-and-yellow combination doesn't. Even eighteen-month-olds immediately figure out the general principle that the two objects have to be the same to make it go, and they generalize that principle to new examples: For instance, they will choose two objects that have the same *shape* to make the machine work. In other experiments, we've shown that children can even figure out that some hidden invisible property makes the machine go, or that the machine works on some abstract logical principle.[38]

---

[38] A. Gopnik, T. Griffiths & C. Lucas, "When younger learners can be better (or at least more open-minded) than older ones," *Curr. Dir. Psychol. Sci.,* 24:2, 87-92 (2015).

You can show this in children's everyday learning, too. Young children rapidly learn abstract intuitive theories of biology, physics, and psychology in much the way adult scientists do, even with relatively little data.

The remarkable machine-learning accomplishments of the recent AI systems, both bottom-up and top-down, take place in a narrow and well-defined space of hypotheses and concepts—a precise set of game pieces and moves, a predetermined set of images. In contrast, children and scientists alike sometimes change their concepts in radical ways, performing paradigm shifts rather than simply tweaking the concepts they already have.

Four-year-olds can immediately recognize cats and understand words, but they can also make creative and surprising new inferences that go far beyond their experience. My own grandson recently explained, for example, that if an adult wants to become a child again, he should try not eating any healthy vegetables, since healthy vegetables make a child grow into an adult. This kind of hypothesis, a plausible one that no grown-up would ever entertain, is characteristic of young children. In fact, my colleagues and I have shown systematically that preschoolers are better at coming up with unlikely hypotheses than older children and adults.[39] We have almost no idea how this kind of creative learning and innovation is possible.

Looking at what children do, though, may give programmers useful hints about directions for computer learning. Two features of children's learning are especially striking. Children are active learners; they don't just passively soak up data like AIs do. Just as scientists experiment, children are intrinsically motivated to extract information from the world around them through their endless play and exploration. Recent studies show that this exploration is more systematic than it looks and is well-adapted to find persuasive evidence to support hypothesis formation and theory choice.[40] Building curiosity into machines and allowing them to actively interact with the world might be a route to more realistic and wide-ranging learning.

Second, children, unlike existing AIs, are social and cultural learners. Humans don't learn in isolation but avail themselves of the accumulated wisdom of past generations. Recent studies show that even preschoolers learn through imitation and by listening to the testimony of others. But they don't simply passively obey their teachers. Instead they take in information from others in a remarkably subtle and sensitive way, making complex inferences about where the information comes from and how trustworthy it is and systematically integrating their own experiences with what they are hearing.[41]

"Artificial intelligence" and "machine learning" sound scary. And in some ways they are. These systems are being used to control weapons, for example, and we really should be scared about that. Still, natural stupidity can wreak far more havoc than artificial intelligence; we humans will need to be much smarter than we have been in the past to properly regulate the new technologies. But there is not much basis for either the apocalyptic or the utopian visions of AIs replacing humans. Until we solve the basic paradox of learning, the best artificial intelligences will be unable to compete with the average human four-year-old.

---

[39] A. Gopnik, *et al.,* "Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood," *Proc. Nat. Acad. Sci.,* 114:30, 7892-99 (2017).

[40] L. Schulz, "The origins of Inquiry: Inductive inference and exploration in early childhood," *Trends Cog. Sci.,* 16:7, 382-89 (2012).

[41] A. Gopnik, *The Gardener and the Carpenter* (New York: Farrar, Straus & Giroux, 2016), chaps. 4 and 5.

Peter Galison's *focus as a science historian is—speaking roughly—on the intersection of theory with experiment.*

"For quite a number of years I have been guided in my work by the odd confrontation of abstract ideas and extremely concrete objects," he once told me, in explaining how he thinks about what he does. At the Washington, Connecticut, meeting he discussed the Cold War tension between engineers (like Wiener) and the administrators of the Manhattan Project (like Oppenheimer: "When [Wiener] warns about the dangers of cybernetics, in part he's trying to compete against the kind of portentous language that people like Oppenheimer [used]: 'When I saw the explosion at Trinity, I thought of the Bhagavad Gita—I am death, destroyer of worlds.' That sense, that physics could stand and speak to the nature of the universe and airforce policy, was repellent and seductive. In a way, you can see that over and over again in the last decades—nanosciences, recombinant DNA, cybernetics: 'I stand reporting to you on the science that has the promise of salvation and the danger of annihilation—and you should pay attention, because this could kill you.' It's a very seductive narrative, and it's repeated in artificial intelligence and robotics."

As a twenty-four-year old, when I first encountered Wiener's ideas and met his colleagues at the MIT meeting I describe in the book's Introduction, I was hardly interested in Wiener's warnings or admonitions. What drove my curiosity was the stark, radical nature of his view of life, based on the mathematical theory of communications in which the message was nonlinear: According to Wiener, "new concepts of communication and control involved a new interpretation of man, of man's knowledge of the universe, and of society." And that led to my first book, which took information theory—the mathematical theory of communications—as a model for all human experience.

In a recent conversation, Peter told me he was beginning to write a book—about building, crashing, and thinking—that considers the black-box nature of cybernetics and how it represents what he thinks of as "the fundamental transformation of learning, machine learning, cybernetics, and the self."

ALGORISTS DREAM OF OBJECTIVITY
Peter Galison

***Peter Galison** is a science historian, Joseph Pellegrino University Professor and co-founder of the Black Hole Initiative at Harvard University, and the author of* Einstein's Clocks and Poincaré's Maps: Empires of Time.

In his second-best book, the great medieval mathematician al-Khwarizmi described the new place-based Indian form of arithmetic. His name, soon sonically linked to "algorismus" (in late medieval Latin) came to designate procedures acting upon numbers—eventually wending its way through "algorithm," (on the model of "logarithm"), into French and on into English. But I like the idea of a modern algorist, even if my spellcheck does not. I mean by it someone profoundly suspicious of the intervention of human judgment, someone who takes that judgment to violate the fundamental norms of what it is to be objective (and therefore scientific).

Near the end of the 20th century, a paper by two University of Minnesota psychologists summarized a vast literature that had long roiled the waters of prediction. One side, they judged, had for all too long held resolutely—and ultimately unethically—to the "clinical method" of prediction, which prized all that was *subjective*: "informal," "in-the-head," and "impressionistic." These clinicians were people (so said the psychologists) who thought they could study their subjects with meticulous care, gather in committees, and make judgment-based predictions about criminal recidivism, college success, medical outcomes, and the like. The other side, the psychologists continued, embodied everything the clinicians did not, embracing the *objective*: "formal," "mechanical," "algorithmic." This the authors took to stand at the root of the whole triumph of post-Galilean science. Not only did science benefit from the actuarial; to a great extent, science *was* the mechanical-actuarial. Breezing through 136 studies of predictions, across domains from sentencing to psychiatry, the authors showed that in 128 of them, predictions using actuarial tables, a multiple-regression equation, or an algorithmic judgment equalled or exceeded in accuracy those using the subjective approach.

They went on to catalog seventeen fallacious justifications for clinging to the clinical. There were the self-interested foot-draggers who feared losing their jobs to machines. Others lacked the education to follow statistical arguments. One group mistrusted the formalization of mathematics; another excoriated what they took to be the actuarial "dehumanizing;" yet others said that the aim was to understand, not to predict. But whatever the motivations, the review concluded that it was downright immoral to withhold the power of the objective over the subjective, the algorithmic over expert judgment.[42]

---

[42] William M. Grove & Paul E. Meehl, "Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The Clinical-Statistical Controversy," *Psychology, Public Policy, and Law*, 2:2, 293-323 (1996).

The algorist view has gained strength. Anne Milgram served as Attorney General of the State of New Jersey from 2007 to 2010. When she took office, she wanted to know who the state was arresting, charging, and jailing, and for what crimes. At the time, she reports in a later TED Talk, she could find almost no data or analytics. By imposing statistical prediction, she continues, law enforcement in Camden during her tenure was able to reduce murders by 41 percent, saving thirty-seven lives, while dropping the total crime rate by 26 percent. After joining the Arnold Foundation as its vice president for criminal justice, she established a team of data scientists and statisticians to create a risk-assessment tool; fundamentally, she construed the team's mission as deciding how to put "dangerous people" in jail while releasing the non-dangerous. "The reason for this," Milgram contended, "is the way we make decisions. Judges have the best intentions when they make these decisions about risk, but they're making them subjectively. They're like the baseball scouts twenty years ago who were using their instinct and their experience to try to decide what risk someone poses. They're being subjective, and we know what happens with subjective decision making, which is that we are often wrong." Her team established nine-hundred-plus risk factors, of which nine were most predictive. The questions, the most urgent questions, for the team were: Will a person commit a new crime? Will that person commit a violent act? Will someone come back to court? We need, concluded Milgram, an "objective measure of risk" that should be inflected by judges' judgment. We know the algorithmic statistical process works. That, she says, is "why Google is Google" and why moneyball wins games.[43]

Algorists have triumphed. We have grown accustomed to the idea that protocols and data can and should guide us in everyday action, from reminders about where we probably want to go next, to the likely occurrence of crime. By now, according to the literature, the legal, ethical, formal, and economic dimensions of algorithms are all quasi-infinite. I'd like to focus on one particular siren song of the algorithm: its promise of objectivity.

Scientific objectivity has a history. That might seem surprising. Isn't the notion—expressed above by the Minnesota psychologists—right? Isn't objectivity co-extensive with science itself? Here it's worth stepping back to reflect on all the epistemic virtues we might value in scientific work. Quantification seems like a good thing to have; so, too, do prediction, explanation, unification, precision, accuracy, certainty, and pedagogical utility. In the best of all possible worlds these epistemic virtues would all pull in the same direction. But they do not—not any more than our ethical virtues necessarily coincide. Rewarding people according to their need may very well conflict with rewarding people according to their ability. Equality, fairness, meritocracy—ethics, in a sense, is all about the adjudication of conflicting goods. Too often we forget that this conflict exists in science, too. Design an instrument to be as sensitive as possible and it often fluctuates wildly, making repetition of a measurement impossible.

"Scientific objectivity" entered both the practice and the nomenclature of science after the first third of the 19th century. One sees this clearly in the scientific atlases that provided scientists with the basic objects of their specialty: There were (and are) atlases of the hand, atlases of the skull, atlases of clouds, crystals, flowers, bubble-chamber pictures, nuclear emulsions, and diseases of the eye. In the 18th century, it was obvious that you would not depict this particular, sun-scorched, caterpillar-chewed clover found outside your house in an atlas. No, you aimed—if you were a genius natural philosopher like Goethe, Albinus, or Cheselden—to observe nature but then to perfect the object in question, to abstract it visually to the ideal. Take a skeleton, view it through a camera lucida, draw it with care. Then correct the "imperfections." The advantage of this parting of the curtains of mere experience was clear: It provided a universal guide, one not attached to the vagaries of individual variation.

---

[43] TED Talk, January 2014, https://www.ted.com/speakers/anne_milgram.

As the sciences grew in scope, and scientists grew in number, the downside of idealization became clearer. It was one thing to have Goethe depict the "ur-plant" or "ur-insect." It was quite another to have a myriad of different scientists each fixing their images in different and sometimes contradictory ways. Gradually, from around the 1830s forward, one begins to see something new: a claim that the image making was done with a minimum of human intervention, that protocols were followed. This could mean tracing a leaf with a pencil or pressing it into ink that was transferred to the page. It meant, too, that one suddenly was proud of depicting the view through a microscope of a natural object even with its imperfections. This was a radical idea: snowflakes shown without perfect hexagonal symmetry, color distortion near the edge of a microscope lens, tissue torn around the edges in the process of its preparation.

Scientific objectivity came to mean that our representations of things were executed by holding back from intervention—even if it meant reproducing the yellow color near the edge of the image under the microscope, despite the fact that the scientist *knew* that the discoloration was from the lens, not a feature of the object of inquiry. The advantage of objectivity was clear: It superseded the desire to see a theory realized or a generally accepted view confirmed. But objectivity came at a cost. You lost that precise, easily teachable, colored, full depth-of-field, artist's rendition of a dissected corpse. You got a blurry, bad depth-of-field, black-and-white photograph that no medical student (nor even many medical colleagues) could use to learn and compare cases. Still, for a long stretch of the 19th century, the virtue of hands-off, self-restraining objectivity was on the rise.

Starting in the 1930s, the hardline scientific objectivity in scientific representation began running into trouble. In cataloging stellar spectra, for example, no algorithm could compete with highly trained observers who could sort them with far greater accuracy and replicability than any purely rule-following procedure. By the late 1940s, doctors had begun learning how to read electroencephalograms. Expert judgment was needed to sort out different kinds of seizure readings, while none of the early attempts to use frequency analysis could match that judgment. Solar magnetograms—mapping the magnetic fields across the sun—required the trained expert to pry the real signal from artifacts that emerged from the measuring instruments. Even particle physicists recognized that they could not program a computer to sort certain kinds of tracks into the right bins; judgment, trained judgment, was needed.

There should be no confusion here: This was not a return to the invoked genius of an 18th-century idealizer. No one thought you could train to be a Goethe who alone among scientists could pick out the universal, ideal form of a plant, insect, or cloud. Expertise could be learned—you could take a course to learn to make expert judgments about electroencephalograms, stellar spectra, or bubble-chamber tracks; alas, no one has ever thought you could take a course that would lead to the mastery of exceptional insight. There can be no royal road to becoming Goethe. In scientific atlas after scientific atlas, one sees explicit argument that "subjective" factors had to be part of the scientific work needed to create, classify, and interpret scientific images.

What we see in so many of the algorists' claims is a tremendous desire to find scientific objectivity precisely by abandoning judgment and relying on mechanical procedures—in the name of scientific objectivity. Many American states have legislated the use of sentencing and parole algorithms. Better a machine, it is argued, than the vagaries of a judge's judgment.

So here is a warning from the sciences. Hands-off algorithmic proceduralism did indeed have its heyday in the 19th century, and of course still plays a role in many of the most successful technical and scientific endeavors. But the idea that mechanical objectivity, construed as binding self-restraint, follows a simple, monotonic curve increasing from the bad impressionistic clinician to the good externalized actuary simply does not answer to the more interesting and nuanced history of the sciences.

There is a more important lesson from the sciences. Mechanical objectivity is a scientific virtue among others, and the hard sciences learned that lesson often. We must do the same in the legal and social scientific domains. What happens, for example, when the secret, proprietary algorithm sends one person to prison for ten years and another for five years, for the same crime? Rebecca Wexler, visiting fellow at the Yale Law School Information Society Project, has explored that question, and the tremendous cost that trade-secret algorithms impose on the possibility of a fair legal defense.[44] Indeed, for a variety of reasons, law enforcement may not want to share the algorithms used to make DNA, chemical, or fingerprint identifications, which puts the defense in a much weakened position to make its case. In the courtroom, objectivity, trade secrets, and judicial transparency may pull in opposite directions. It reminds me of a moment in the history of physics. Just after World War II, the film giants Kodak and Ilford perfected a film that could be used to reveal the interactions and decays of elementary particles. The physicists were thrilled, of course—until the film companies told them that the composition of the film was a trade secret, so the scientists would never gain complete confidence that they understood the processes they were studying. Proving things with unopenable black boxes can be a dangerous game for scientists, and doubly so for criminal justice.

Other critics have underscored how perilous it is to rely on an accused (or convicted) person's address or other variables that can easily become, inside the black box of algorithmic sentencing, a proxy for race. By dint of everyday experience, we have grown used to the fact that airport security is different for children under the age of twelve and adults over the age of seventy-five. What factors do we want the algorists to have in their often hidden procedures? Education? Income? Employment history? What one has read, watched, visited, or bought? Prior contact with law enforcement? How do we want algorists to weight those factors? Predictive analytics predicated on mechanical objectivity comes at a price. Sometimes it may be a price worth paying; sometimes that price would be devastating for the just society we want to have.

More generally, as the convergence of algorithms and Big Data governs a greater and greater part of our lives, it would be well worth keeping in mind these two lessons from the history of the sciences: Judgment is not the discarded husk of a now pure objectivity of self-restraint. And mechanical objectivity is a virtue competing among others, not the defining essence of the scientific enterprise. They are lessons to bear in mind, even if algorists dream of objectivity.

---

[44] Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System," 70 *Stanford Law Review,* XXX (2018).

In the past decade, genetic engineering has caught up with computer science with regard to how new scientific initiatives are shaping our lives. Genetic engineer **George Church***, a pioneer of the revolution in reading and writing biology, is central to this new landscape of ideas. He thinks of the body as an operating system, with engineers taking the place of traditional biologists in retooling stripped-down components of organisms (from atoms to organs) in much the same vein as in the late 1970s, when electrical engineers were working their way to the first personal computer by assembling circuit boards, hard drives, monitors, etc. George created and is director of the Personal Genome Project, which provides the world's only open-access information on human genomic, environmental, and trait data (GET) and sparked the growing DNA ancestry industry.*

He was instrumental in laying the groundwork for President Obama's 2013 BRAIN (Brain Research through Advancing Innovative Neurotechnologies) Initiative—in aid of improving the brains of human beings to the point where, for much of what sustains us, we might not need the help of (potentially dicey) AIs. "It could be that some of the BRAIN Initiative projects allow us to build human brains that are more consistent with our ethics and capable of doing advanced tasks like artificial intelligence," George has said. "The safest path by far is getting humans to do all the tasks that they would like to delegate to machines, but we're not yet firmly on that super-safe path."

More recently, his crucially important pioneering use of the enzyme CRISPR (as well as methods better than CRISPR) to edit the genes of human cells is sometimes missed by the media in the telling of the CRISPR origins story.

George's attitude toward future forms of artificial general intelligence is friendly, as evinced in the essay that follows. At the same time, he never loses sight of the AI-safety issue. On that subject, he recently remarked: "The main risk in AI, to my mind, is not so much whether we can mathematically understand what they're thinking; it's whether we're capable of teaching them ethical behavior. We're barely capable of teaching each other ethical behavior."

THE RIGHTS OF MACHINES
George M. Church

***George M. Church** is Robert Winthrop Professor of Genetics at Harvard Medical School; Professor of Health Sciences and Technology, Harvard-MIT; and co-author (with Ed Regis) of* Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves.

In 1950, Norbert Wiener's *The Human Use of Human Beings* was at the cutting edge of vision and speculation in proclaiming that

> the machine like the djinnee, which can learn and can make decisions on the basis of its learning, will in no way be obliged to make such decisions as we should have made, or will be acceptable to us. . . . Whether we entrust our decisions to machines of metal, or to those machines of flesh and blood which are bureaus and vast laboratories and armies and corporations, . . . [t]he hour is very late, and the choice of good and evil knocks at our door.

But this was his book's denouement, and it has left us hanging now for sixty-eight years, lacking not only prescriptions and proscriptions but even a well-articulated "problem statement." We have since seen similar warnings about the threat of our machines, even in the form of outreach to the masses, via films like *Colossus: The Forbin Project* (1970), *The Terminator* (1984), *The Matrix* (1999), and *Ex Machina* (2015). But now the time is ripe for a major update, with fresh, new perspectives—notably focused on generalizations of our "human" rights and our existential needs.

Concern has tended to focus on "us versus them [robots]" or "grey goo [nanotech]" or "monocultures of clones [bio]." To extrapolate current trends: What if we could make or grow almost anything and engineer any level of safety and efficacy desired? Any thinking being (made of any arrangement of atoms) could have access to any technology.

Probably we should be less concerned about us-versus-them and more concerned about the rights of all sentients in the face of an emerging unprecedented diversity of minds. We should be harnessing this diversity to minimize global existential risks, like supervolcanoes and asteroids.

But should we say "should"? (Disclaimer: In this and many other cases, when a technologist describes a societal path that "could," "would," or "should" happen, this doesn't necessarily equate to the preferences of the author. It could reflect warning, uncertainty, and/or detached assessment.) Roboticist Gianmarco Veruggio and others have raised issues of roboethics since 2002; the U.K. Department of Trade and Industry and the RAND spin-off Institute for the Future have raised issues of robot rights since 2006.

"Is versus ought"

It is commonplace to say that science concerns "is," not "ought." Stephen Jay Gould's "non-overlapping magisteria" view argues that facts must be completely distinct from values. Similarly, the 1999 document *Science and Creationism* from the U.S. National Academy of Sciences noted that "science and religion occupy two separate realms." This division has been critiqued by evolutionary biologist Richard Dawkins, myself, and others. We can discuss "should" if framed as "we should do X in order to achieve Y." Which Y should be a high priority is not necessarily settled by democratic vote but might be settled by Darwinian vote. Value systems and religions wax and wane, diversify, diverge, and merge just as living species do: subject to selection. The ultimate "value" (the "should") is survival of genes and memes.

Few religions say that there is no connection between our physical being and the spiritual world. Miracles are documented. Conflicts between Church doctrine and Galileo and Darwin are eventually resolved. Faith and ethics are widespread in our species and can be studied using scientific methods, including but not limited to fMRI, psychoactive drugs, questionnaires, et cetera.

Very practically, we have to address the ethical rules that should be built in, learned, or probabilistically chosen for increasingly intelligent and diverse machines. We have a whole series of trolley problems. At what number of people in line for death should the computer decide to shift a moving trolley to one person? Ultimately this might be a deep-learning problem—one in which huge databases of facts and contingencies can be taken into account, some seemingly far from the ethics at hand.

For example, the computer might infer that the person who would escape death if the trolley is left alone is a convicted terrorist recidivist loaded up with doomsday pathogens, or a saintly POTUS—or part of a much more elaborate chain of events in detailed alternative realities. If one of these problem descriptions seems paradoxical or illogical, it may be that the authors of the trolley problem have adjusted the weights on each sides of the balance such that hesitant indecision is inevitable.

Alternatively, one can use misdirection to rig the system, such that the error modes are not at the level of attention. For example, in the Trolley Problem, the real ethical decision was made years earlier when pedestrians were given access to the rails—or even before that, when we voted to spend more on entertainment than on public safety. Questions that at first seem alien and troubling, like "Who owns the new minds, and who pays for their mistakes?" are similar to well-established laws about who owns and pays for the sins of a corporation.

The Slippery Slopes

We can (over)simplify ethics by claiming that certain scenarios won't happen. The technical challenges or the bright red lines that cannot be crossed are reassuring, but the reality is that once the benefits seem to outweigh the risks (even briefly and barely), the red lines shift. Just before Louise Brown's birth in 1978, many people were worried that she "would turn out to be a little monster, in some way, shape or form, deformed, something wrong with her."[45] Few would hold this view of in-vitro fertilization today.

What technologies are lubricating the slope toward multiplex sentience? It is not merely deep machine-learning algorithms with Big Iron. We have engineered rodents to be significantly better at a variety of cognitive tasks as well as to exhibit other relevant traits, such as persistence and low anxiety. Will this be applicable to animals that are already at the door of humanlike intelligence? Several show self-recognition in a mirror test—chimpanzees, bonobos, orangutans, some dolphins and whales, and magpies.

---

[45] "Then, Doctors 'All Anxious' About Test-tube Baby" http://edition.cnn.com/2003/HEALTH/parenting/07/25/cnna.copperman/

Even the bright red line for human manipulation of human beings shows many signs of moving or breaking completely. More than 2,300 approved clinical trials for gene therapy are in progress worldwide. A major medical goal is the treatment or prevention of cognitive decline, especially in light of our rapidly aging global demographic. Some treatments of cognitive decline will include cognitive enhancements (drugs, genes, cells, transplants, implants, and so on). These will be used off-label. The rules of athletic competition (e.g., banning augmentation with steroids or erythropoietin) do not apply to intellectual competition in the real world. Every bit of progress on cognitive decline is in play for off-label use.

Another frontier of the human use of humans is "brain organoids." We can now accelerate developmental biology. Processes that normally take months can happen in four days in the lab using the right recipes of transcription factors. We can make brains that, with increasing fidelity, recapitulate the differences between people born with aberrant cognitive abilities (e.g., microcephaly). Proper vasculature (veins, arteries, and capillaries) missing from earlier successes are now added, enabling brain organoids to surpass the former sub-microliter limit to possibly exceed the 1.2-liter size of modern human brains (or even the 5-liter elephant or 8-liter sperm whale brains).

Conventional Computers versus Bio-electronic Hybrids

As Moore's Law miniaturization approaches its next speed bump (surely not a solid wall), we see the limits of the stochastics of dopant atoms in silicon slabs and the limits of beam-fabrication methods at around 10-nanometer feature size. Power (energy consumption) issues are also apparent: The great Watson, winner of *Jeopardy!*, used 85,000 watts real time, while the human brains were using 20 watts each. To be fair, the human body needs 100 watts to operate and twenty years to build, hence about 6 trillion joules of energy to "manufacture" a mature human brain. The cost of manufacturing Watson-scale computing is similar. So why aren't humans displacing computers?

For one, the *Jeopardy!* contestants' brains were doing far more than information retrieval—much of which would be considered mere distractions by Watson (e.g., cerebellar control of smiling). Other parts allow leaping out of the box with transcendence unfathomable by Watson, such as what we see in Einstein's five *annus mirabilis* papers of 1905. Also, humans consume more energy than the minimum (100 W) required for life and reproduction. People in India use an average of 700 W per person; it's 10,000 W in the U.S. Both are still less than the 85,000 watts Watson uses. Computers can become more like us via neuromorphic computing, possibly a thousandfold. But human brains could get more efficient, too. The organoid brain-in-a-bottle could get closer to the 20 W limit. The idiosyncratic advantages of computers for math, storage, and search, faculties of limited use to our ancestors, could be designed and evolved anew in labs.

Facebook, the National Security Agency, and others are constructing exabyte-scale storage facilities at more than a megawatt and four hectares, while DNA can store that amount in a milligram. Clearly, DNA is not a mature storage technology, but with Microsoft and Technicolor doubling down on it, we would be wise to pay attention. The main reason for the 6 trillion joules of energy required to get a productive human mind is the twenty years required for training.

Even though a supercomputer can "train" a clone of zemself in seconds, the energy cost of producing a mature silicon clone is comparable. Engineering (*Homo*) prodigies might make a small impact on this slow process, but speeding up development and implanting extensive memory (as DNA-exabytes or other means) could reduce duplication time of a bio-computer to close to the doubling time of cells (ranging from eleven minutes to twenty-four hours). The point is that while we may not know what ratio of bio/homo/nano/robo hybrids will be dominant at each step of our accelerating evolution, we can aim for high levels of humane, fair, and safe treatment ("use") of one another.

Bills of Rights date back to 1689 in England. FDR proclaimed the "Four Freedoms"—freedom of speech, freedom of conscience, freedom from fear, and freedom from want. The U.N.'s Universal Declaration of Human Rights in 1948 included the right to life; the prohibition of slavery; defense of rights when violated; freedom of movement; freedom of association, thought, conscience, and religion; social, economic, and cultural rights; duties of the individual to society; and prohibition of use of rights in contravention of the purposes and principles of the United Nations.

The "universal" nature of these rights is not universally embraced and is subject to extensive critique and noncompliance. How does the emergence of non-*Homo*-intelligences affect this discussion? At a minimum, it is becoming rapidly difficult to hide behind vague intuition for ethical decisions—"I know it when I see it" (U.S. Supreme Court Justice Potter Stewart, 1964) or the "wisdom of repugnance" **(**aka "yuck factor," Leon Kass, 1997), or vague appeals to "common sense." As we have to deal with minds alien to us, sometimes quite literal from our viewpoint, we need to be explicit—yea, even algorithmic.

Self-driving cars, drones, stock-market transactions, NSA searches, et cetera, require rapid, pre-approved decision making. We may gain insights into many aspects of ethics that we have been trying to pin down and explain for centuries. The challenges have included conflicting priorities, as well as engrained biological, sociological, and semi-logical cognitive biases. Notably far from consensus in universal dogmas about human rights are notions of privacy and dignity, even though these influence many laws and guidelines.

Humans might want the right to march in to read (and change) the minds of computers to see why they're making decisions at odds with our (*Homo*) instincts. Is it not fair for machines to ask the same of us? We note the growth of movements toward transparency in potential financial conflicts; "open-source" software, hardware, and wetware; the Fair Access to Science and Technology Research Act (FASTR); and the Open Humans Foundation.

In his 1976 book *Computer Power and Human Reason,* Joseph Weizenbaum argued that machines should not replace *Homo* in situations requiring respect, dignity, or care, while others (author Pamela McCorduck and computer scientists like John McCarthy and Bill Hibbard) replied that machines can be more impartial, calm, and consistent and less abusive or mischievous than people in such positions.

Equality

What did the thirty-three-year-old Thomas Jefferson mean in 1776 when he wrote, "We hold these Truths to be self-evident, that all Men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty, and the Pursuit of Happiness"? The spectrum of current humans is vast. In 1776, "Men" did not include people of color or women. Even today, humans born with congenital cognitive or behavioral issues are destined for unequal (albeit in most cases compassionate) treatment—Down syndrome, Tay-Sachs disease, Fragile X syndrome, cerebral palsy, and so on.

And as we change geographical location and mature, our unequal rights change dramatically. Embryos, infants, children, teens, adults, patients, felons, gender identities and gender preferences, the very rich and very poor—all of these face different rights and socioeconomic realities. One path to new mind-types obtaining and retaining rights similar to the most elite humans would be to keep a *Homo* component, like a human shield or figurehead monarch/CEO, signing blindly enormous technical documents, making snap financial, health, diplomatic, military, or security decisions. We will probably have great difficulty pulling the plug, modifying, or erasing (killing) a computer and its memories—especially if it has befriended humans and made spectacularly compelling pleas for survival (as all excellent researchers fighting for their lives would do).

Even Scott Adams, creator of *Dilbert*, has weighed in on this topic, supported by experiments at Eindhoven University in 2005 noting how susceptible humans are to a robot-as-victim equivalent of the Milgram experiments done at Yale beginning in 1961. Given the many rights of corporations, including ownership of property, it seems likely that other machines will obtain similar rights, and it will be a struggle to maintain inequities of selective rights along multi-axis gradients of intellect and ersatz feelings.

Radically Divergent Rules for Humans versus Nonhumans and Hybrids

The divide noted above for intra *Homo sapiens* variation in rights explodes into a riot of inequality as soon as we move to entities that overlap (or will soon) the spectrum of humanity. In Google Street View, people's faces and car license plates are blurred out. Video devices are excluded from many settings, such as courts and committee meetings. Wearable and public cameras with facial-recognition software touch taboos. Should people with hyperthymesia or photographic memories be excluded from those same settings?

Shouldn't people with prosopagnosia (face blindness) or forgetfulness be able to benefit from facial-recognition software and optical character recognition wherever they go, and if them, then why not everyone? If we all have those tools to some extent, shouldn't we all be able to benefit?

These scenarios echo Kurt Vonnegut's 1961 short story "Harrison Bergeron," in which exceptional aptitude is suppressed in deference to the mediocre lowest common denominator of society. Thought experiments like John Searle's Chinese Room and Isaac Asimov's Three Laws of Robotics all appeal to the sorts of intuitions plaguing human brains that Daniel Kahneman, Amos Tversky, and others have demonstrated. The Chinese Room experiment posits that a mind composed of mechanical and *Homo sapiens* parts cannot be conscious, no matter how competent at intelligent human (Chinese) conversation, unless a human can identify the source of the consciousness and "feel" it. Enforced preference for Asimov's First and Second Laws favor human minds over any other mind meekly present in his Third Law, of self-preservation.

If robots don't have exactly the same consciousness as humans, then this is used as an excuse to give them different rights, analogous to arguments that other tribes or races are less than human. Do robots already show free will? Are they already self-conscious? The robots Qbo have passed the "mirror test" for self-recognition and the robots NAO have passed a related test of recognizing their own voice and inferring their internal state of being, mute or not.

For free will, we have algorithms that are neither fully deterministic nor random but aimed at nearly optimal probabilistic decision making. One could argue that this is a practical Darwinian consequence of game theory. For many (not all) games/problems, if we're totally predictable or totally random, then we tend to lose.

What is the appeal of free will anyway? Historically it gave us a way to assign blame in the context of reward and punishment on Earth or in the afterlife. The goals of punishment might include nudging the priorities of the individual to assist the survival of the species. In extreme cases, this could include imprisonment or other restrictions, if Skinnerian positive/negative reinforcement is inadequate to protect society. Clearly, such tools can apply to free will, seen broadly—to any machine whose behavior we'd like to manage.

We could argue as to whether the robot actually experiences subjective qualia for free will or self-consciousness, but the same applies to evaluating a human. How do we know that a sociopath, a coma patient, a person with Williams syndrome, or a baby has the same free will or self-consciousness as our own? And what does it matter, practically? If humans (of any sort) convincingly claim to experience consciousness, pain, faith, happiness, ambition, and/or utility to society, should we deny them rights because their hypothetical qualia are hypothetically different from ours?

The sharp red lines of prohibition, over which we supposedly will never step, increasingly seem to be short-lived and not sensible. The line between human and machines blurs, both because machines become more humanlike and humans become more machine-like—not only since we increasingly blindly follow GPS scripts, reflex tweets, and carefully crafted marketing, but also as we digest ever more insights into our brain and genetic programming mechanisms. The NIH BRAIN Initiative is developing innovative technologies and using these to map out the connections and activity of mental circuitry so as to improve electronic and synthetic neurobiological ware.

Various red lines depend on genetic exceptionalism, in which genetics is considered permanently heritable (although it is provably reversible), whereas exempt (and lethal) technologies, like cars, are for all intents and purposes irreversible due to social and economic forces. Within genetics, a red line makes us ban or avoid genetically modified foods but embrace genetically modified bacteria making insulin, or genetically modified humans—witness mitochondrial therapies approved in Europe for human adults and embryos.

The line for germline manipulation seems less sensible than the usual, practical line drawn at safety and efficacy. Marriages of two healthy carriers of the same genetic disease have a choice between no child of their own, 25-percent loss of embryos via abortion (spontaneous or induced), 80-percent loss via in-vitro fertilization, or potential zero-percent embryo loss via sperm (germline) engineering. It seems premature to declare this last option unlikely.

For “human subject research,” we refer to the 1964 Declaration of Helsinki, keeping in mind the 1932-1972 Tuskegee syphilis experiment, possibly the most infamous biomedical research study in U.S. history. In 2015, the Nonhuman Rights Project filed a lawsuit with the New York State Supreme Court on behalf of two chimpanzees kept for research by Stony Brook University. The appellate court decision was that chimps are not to be treated as legal persons since they “do not have duties and responsibilities in society,” despite Jane Goodall’s and others’ claim that they do, and despite arguments that such a decision could be applied to children and the disabled.[46]

What prevents extension to other animals, organoids, machines, and hybrids? As we (e.g., Hawking, Musk, Tallinn, Wilczek, Tegmark) have promoted bans on “autonomous weapons,” we have demonized one type of “dumb” machine, while other machines—for instance, those composed of many *Homo sapiens* voting—can be more lethal and more misguided.

Do transhumans roam the Earth already? Consider the “uncontacted peoples,” such as the Sentinelese and Andamanese of India, the Korowai of Indonesia, the Mashco-Piro of Peru, the Pintupi of Australia, the Surma of Ethiopia, the Ruc of Vietnam, the Ayoreo-Totobiegosode of Paraguay, the Himba of Namibia, and dozens of tribes in Papua New Guinea. How would they or our ancestors respond? We could define “transhuman” as people and culture not comprehensible to humans living in a modern, yet un-technological culture.

Such modern Stone Age people would have great trouble understanding why we celebrate the recent LIGO gravity-wave evidence supporting the hundred-year-old general theory of relativity. They would scratch their heads as to why we have atomic clocks, or GPS satellites so we can find our way home, or why and how we have expanded our vision from a narrow optical band to the full spectrum from radio to gamma. We can move faster than any other living species; indeed, we can reach escape velocity from Earth and survive in the very cold vacuum of space.

If those characteristics (and hundreds more) don’t constitute transhumanism, then what would? If we feel that the judge of transhumanism should not be fully paleo-culture humans but recent humans, then how would we ever reach transhuman status? We “recent humans” may always be capable of comprehending each new technological increment—never adequately surprised to declare arrival at a (moving) transhuman target. The science-fiction prophet William Gibson said, “The future is already here—it’s just not very evenly distributed.” While this underestimates the next round of “future,” certainly millions of us are transhuman already—with most of us asking for more. The question “What was a human?” has already transmogrified into “What were the many kinds of transhumans?. . . And what were their rights?”

---

[46] https://www.nbcnews.com/news/us-news/lawyer-denying-chimpanzees-rights-could-backfire-disabled-n734566.

Caroline A. Jones' *interest in modern and contemporary art is enriched by a willingness to delve into the technologies involved in its production, distribution, and reception. "As an art historian, a lot of my questions are about what kind of art we can make, what kind of thought we can make, what kind of ideas we can make that could stretch the human beyond our stubborn, selfish, 'only concerned with our small group' parameters. The philosophers and philosophies I'm drawn to are those that question the Western obsession with individualism. Those are coming from so many different places, and they're reviving so many different kinds of questions and problems that were raised in the 1960s."*

She has recently turned her attention to the history of cybernetics. Her MIT course, "Automata, Automatism, Systems, Cybernetics," explores the history of the human/machine interface in terms of feedback, exploring the cultural rather than engineering uptake of this idea. She begins with primary readings by Wiener, Shannon, and Turing and then pivots from the scientists and engineers to the work and ideas of artists, feminists, postmodern theorists. Her goal: to come up with a new central paradigm of evolution that's culture-based—"communalism and interspecies symbiosis rather than survival of the fittest."

As a historian, Caroline draws a distinction between what she has termed "left cybernetics" and "right cybernetics": "What do I mean by left cybernetics? In one sense, it's a pun or a joke: the cybernetics that was 'left' behind. On another level, it's a vague political grouping connoting our Left Coast: California, Esalen, the group that Dave Kaiser calls the 'hippie physicists.' It's not an adequate term, but it's a way of recognizing that there was a group beholden to the military-industrial complex, sometimes very unhappily, who gave us the tools to critique it."

THE ARTISTIC USE OF CYBERNETIC BEINGS

Caroline A. Jones

***Caroline A. Jones** is a professor of art history in the Department of Architecture at MIT and author of* Eyesight Alone: Clement Greenberg's Modernism and the Bureaucratization of the Senses; Machine in the Studio: Constructing the Postwar American Artist; *and* The Global Work of Art.

> Cybernated art is very important, but art for cybernated life is more important.
> — *Nam June Paik, 1966*

Artificial intelligence was not what artists first wanted out of cybernetics, once Norbert Wiener's *The Human Use of Human Beings: Cybernetics and Society* came out in 1950. The range of artists who identified themselves with cybernetics in the fifties and sixties initially had little access to "thinking machines." Moreover, craft-minded engineers had already been making turtles, jugglers, and light-seeking robot babes, not giant brains. Using breadboards, copper wire, simple switches, and electronic sensors, artists followed cyberneticians in making sculptures and environments that simulated interactive sentience—analog movements and interfaces that had more to do with instinctive drives and postwar sexual politics than the automation of knowledge production. Now obscured by an ideology of a free-floating "intelligence" untethered by either hardware or flesh, AI has forgotten the early days of cybernetics' uptake by artists. Those efforts are worth revisiting; they modeled relations with what the French philosophers Gilles Deleuze and Félix Guattari have called the "machinic phylum," having to do with how humans think and feel in bodies engaged with a physical, material, emotionally stimulating, and signaling world.

Cybernetics now seems to have collapsed into an all-pervasive discourse of AI that was far from preordained. "Cybernetics," as a word, claimed postwar newness for concepts that were easily four centuries old: notions of feedback, machine damping, biological homeostasis, logical calculation, and systems thinking that had been around since the Enlightenment (boosted by the Industrial Revolution). The names in this lineage include Descartes, Leibniz, Sadi Carnot, Clausius, Maxwell, and Watt. Wiener's coinage nonetheless had profound cultural effects.[47] The ubiquity today of the prefix "cyber-" confirms the desire for a crisp signifier of the tangled relations between humans and machines. In Wiener's usage, things "cyber" simply involved "control and communication in the animal and the machine." But after the digital revolution, "cyber" moved beyond servomechanisms, feedback loops, and switches to encompass software, algorithms, and cyborgs. The work of cybernetically inclined artists concerns the emergent behaviors of life that elude AI in its current condition.

As to that original coinage, Wiener had reached back to the ancient Greek to borrow the word for "steersman" (κυβερνήτης / *kubernétés*), a masculine figure channeling power and instinct at the helm of a ship, who read the waves, judged the wind, kept a hand on the tiller, and directed the slaves as they mindlessly (mechanically) churned their oars. The Greek had already migrated into modern English via Latin, going from *kuber-* to *guber*—the root of "gubernatorial" and "governor," another term for masculine control, deployed by James Watt to describe his 19th-century device for modulating a runaway steam engine. Cybernetics thus took ideas that had long analogized people and devices and generalized them to an applied science by adding that "-ics." Wiener's three c's (command, control, communication) drew on the mathematics of probability to formalize systems (whether biological or mechanical) theorized as a set of inputs of information achieving outputs of actions in an environment—a muscular, fleshy agenda often minimized in genealogies of AI.

---

[47] Wiener later had to admit the earlier coinage of the word in 1834 by André-Marie Ampère, who had intended it to mean the "science of government," a concept that remained dormant until the 20th century.

But the etymology does little to capture the excitement felt by participants, as mathematics joined theoretical biology (Arturo Rosenblueth) and information theory (Claude Shannon, Walter Pitts, Warren McCulloch) to produce a barrage of interdisciplinary research and publications viewed as changing not just the way science was done but the way future humans would engage with the technosphere. As Wiener put it, "We have modified our environment so radically that we must now modify ourselves in order to exist."[48] The pressing question is: How are we modifying ourselves? Are we going in the right direction or have we lost our way, becoming the tools of our tools? Revisiting the early history of humanist/artists' contribution to cybernetics may help direct us toward a less perilous, more ethical future.

The year 1968 was a high-water mark of the cultural diffusion and artistic uptake of the term. In that year, the Howard Wise gallery opened its show of Wen-Ying Tsai's "Cybernetic Sculpture" in midtown Manhattan, and Polish émigré Jasia Reichardt opened her exhibition "Cybernetic Serendipity" at London's ICA. (The "Cybernetic" in her title was intended to evoke "made by or with computers," even though most of the artworks on view had no computers, as such, in their responsive circuits.) The two decades between 1948 and 1968 had seen both the fanning out of cybernetic concepts into a broader culture and the spread of computation machines themselves in a slow migration from proprietary military equipment, through the multinational corporation, to the academic lab, where access began to be granted to artists. The availability of cybernetic components—"sensor organs" (electronic eyes, motion sensors, microphones) and "effector organs" (electronic "breadboards," switches, hydraulics, pneumatics)—on the home hobbyist front rendered the computer less an "electronic brain" than an adjunct organ in a kit of parts. There was not yet a ruling metaphor of "artificial intelligence." So artists were bricoleurs of electronic bodies, interested in actions rather than calculation or cognition. There were inklings of "computer" as calculator in the drive toward *Homo rationalis*, but more in aspiration than achievement.

In light of today's digital convergence in art/science imaging tools, Reichardt's show was prophetic in its insistence on confusing the boundaries between art and what we might dub "creative applied science." According to the catalog, "no visitor to the exhibition, unless he reads all the notes relating to all the works, will know whether he is looking at something made by an artist, engineer, mathematician, or architect." So the comically dysfunctional robot by Nam June Paik, *Robot K-456* (1964), featured on the catalog's cover and described as "a female robot known for her disturbing and idiosyncratic behavior," would face off against a balletic *Colloquy of Mobiles* (1968) from second-order cybernetician Gordon Pask. Pask worked with a London theater designer to craft a spindly "male" apparatus of hinges and rods, set up to communicate with bulbous "female" fiberglass entities nearby. Whether anyone could actually map the quiddities of the program (or glean its reactionary gender theater) without reading the catalog essay is an open question. What is significant is Pask's focus on the behaviors of his automata, their interactivity, their responsiveness within an artificially modulated environment, and their "reflection" of human behaviors.

---

[48] *The Human Use of Human Beings* (1954 edition), p. 46.

The ICA's "Cybernetic Serendipity" introduced an important paradigm: the machinic ecosystem, in which the viewer was a biological part, tasked with figuring out just what the triggers for interaction might be. The visitors in those London galleries suddenly became "cybernetic organisms"—cyborgs—since to experience the art adequately, one needed to enter a kind of symbiotic colloquy with the servomechanisms. This turn toward human-machine interactive environments as an aesthetic becomes clearer when we examine a few other artworks from the period, beginning with one constituting an early instance of emergent behavior—*Senster,* the interactive sculpture by artist/engineer Edward Ihnatowicz (1970), celebrated by medical robotics engineer Alex Zivanovic, editor of a Web site devoted to Ihnatowicz's little-known career, as "one of the first computer controlled interactive robotic works of art." Here, "the computer" makes its entry (albeit a twelve-bit, limited device). But rather than "intelligence," Ihnatowicz sought to make an avatar of affective behavior. Key to *Senster's* uncanny success was the programming with which Ihnatowicz constrained the fifteen-foot-long hydraulic apparatus (its hinge design and looming appearance inspired by a lobster claw) to convey shyness in responding to humans in its proximity. *Senster's* sound channels and motion sensors were set to recoil at loud noises and sudden aggressive movements. Only those humans willing to speak softly and modulate their gestures would be rewarded by *Senster's* quiet, inquisitive approach—an experience that became real for Ihnatowicz himself when he first assembled the program and the machine turned to him solicitously after he'd cleared his throat.

In these artistic uses of cybernetic beings, we sense a growing necessity to train the public to experience itself as embedded in a technologized environment, modifying itself to communicate intuitively with machines. This necessity had already become explicit in Tsai's "Cybernetic Sculpture" show. Those experiencing his immersive installation were expected to experiment with machinic life: What behaviors would trigger the servomechanisms? Likely, the human gallery attendant would have had to explain the protocol: "Clap your hands—that gets the sculptures to respond." As an early critic described it:

> A grove of slender stainless-steel rods rises from a plate. This base vibrates at 30 cycles per second; the rods flex rapidly, in harmonic curves. Set in a dark room, they are lit by strobes. The pulse of the flashing lights varies—they are connected to sound and proximity sensors. The result is that when one approaches a Tsai or makes a noise in its vicinity, the thing responds. The rods appear to move; there is a shimmering, a flashing, an eerie ballet of metal, whose apparent movements range from stillness to jittering and back to a slow, indescribably sensuous undulation.[49]

---

[49] Robert Hughes, *Time* magazine (October 2, 1972) review of Tsai exhibition at Denise René gallery.

Like *Senster,* the apparatus stimulated (and simulated) an affective rather than rational interaction. Humans felt they were encountering behaviors indicative of responsive life; Tsai's entities were often classed as "vegetal" or "aquatic." Such environmental and kinetic ambitions were widespread in the international art world of the time. Beyond the stable at Howard Wise, there were the émigrés forming the collective GRAV in Paris, the "cybernetic architectures" of Nicolas Schöffer, the light and plastic gyrations of the German Zero Gruppe, and so on—all defining and informing the genre of installation art to come.

The artistic use of cybernetic beings in the late sixties made no investment in "intelligence." Knowing machines were dumb and incapable of emotion, these creators were confident in staging frank simulations. What interested them were machinic motions evoking drives, instincts, and affects; they mimicked sexual and animal behaviors, as if below the threshold of consciousness. Such artists were uninterested in the manipulation of data or information (although Hans Haacke would move in that direction by 1972 with his "Real-Time Systems" works). The cybernetic culture that artists and scientists were putting in place on two continents embedded the human in the technosphere and seduced perception with the graceful and responsive behaviors of the machinic phylum. "Artificial" and "natural" intertwined in this early cybernetic aesthetic.

But it wouldn't end here. Crucial to the expansion of this uncritical, largely masculine set of cybernetic environments would be a radical, critical cohort of astonishing women artists emerging in the 1990s, fully aware of their predecessors in art and technology but perhaps more inspired by the feminist founders of the 1970 journal *Radical Software* and the cultural blast of Donna Haraway's inspiring 1984 polemic, "A Cyborg Manifesto." The creaky gender theater of Paik and Pask, the innocent creatures of Ihnatowicz and Tsai, were mobilized as savvy, performative, and postmodern, as in Lynn Hershman Leeson's *Dollie Clone Series* (1995-98) consisting of the interactive assemblages *CyberRoberta* and *Tillie, the Telerobotic Doll,* who worked the technosphere with the professionalism of burlesque, winking and folding us viewers into an explicit consciousness of our voyeuristic position as both seeing subjects and objects-to-be-looked-at.

The "innocent" technosphere established by male cybernetic sculptors of the 1960s was, by the 1990s, identified by feminist artists as an entirely suffusive condition demanding our critical attention. At the same time, feminists tackled the question of whose "intelligence" AI was attempting to simulate. For an artist such as Hershman Leeson, responding to the technical "triumph" of cloning Dolly the sheep, it was crucial to draw the connection between meat production and "meat machines." Hershman Leeson produced "dolls" as clones, offering a critical framing of the way contemporary individuation had become part of an ideological, replicative, plastic realm.

While the technofeminists of the 1990s and into the 2000s weren't all cyber all the time, their works nonetheless complicated the dominant machinic and kinetic qualities of male artists' previous techno-environments. The androgynous tele-cyborg in Judith Barry's *Imagination, Dead Imagine* (1991), for example, had no moving parts: He/she was comprised of pure signals, flickering projections on flat surfaces. In her setup, Barry commented on the alienating effects of late-20th-century technology. The image of an androgynous head fills an enormous cube made of ten-foot-square screens on five sides, mounted on a ten-foot-wide mirrored base. A variety of viscous and unpleasant-looking fluids (yellow, reddish-orange, brown), dry materials (sawdust? flour?), and even insects drizzle or dust their way down the head, whose stoic sublimity is made gorgeously virtual on the work's enormous screens. *Dead Imagine,* through its large-scale and cubic "Platonic" form, remains both artificial and locked into the body—refusing a detached "intelligence" as being no intelligence at all.

Artists in the new millennium inherit this critical tradition and inhabit the current paradigms of AI, which has slid from partial simulations to claims of intelligence. In the 1955 proposal thought to be the first printed usage of the phrase "artificial intelligence," computer scientist John McCarthy and his colleagues Marvin Minsky, Nathaniel Rochester, and Claude Shannon conjectured that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." This modest theoretical goal has inflated over the past sixty-four years and is now expressed by Google DeepMind as an ambition to "Solve intelligence." Crack the code! But unfortunately, what we hear cracking is not code but small-scale capitalism, the social contract, and the scaffolding of civility. Taking away the jobs of taxi and truck drivers, roboticizing direct marketing, hegemonizing entertainment, privatizing utilities, and depersonalizing health care—are these the "whips" that Wiener feared we would learn to love?

Artists can't solve any of this. But they can remind us of the creative potential of the paths not taken—the forks in the road that were emerging around 1970, before "information" became capital and "intelligence" equaled data harvesting. Richly evocative of what can be done with contemporary tools when revisiting earlier possibilities is French artist Philippe Parreno's "firefly piece," so nicknamed to avoid having to iterate its actual title: *With a Rhythmic Instinction to Be Able to Travel Beyond Existing Forces of Life* (2014). Described by the artist as "an automaton," the sculptural installation juxtaposes a flickering projection of black-and-white drawings of fireflies with a band of oscillating green-on-black binary figures. The drawings and binary figures are animated using algorithms from mathematician John Horton Conway's 1970 Game of Life, a "cellular automaton."

Conway set up parameters for any square ("cell") to be lit ("alive") or dark ("dead") in an infinite, two-dimensional grid. The rules are summarized as follows: A single cell will quickly die of loneliness. But a cell touching three or more other "live" cells will also die, "due to crowding." A cell survives and thrives if it has just two neighbors . . . and so on. As one cell dies, it may create the conditions for other cells to survive, yielding patterns that appear to move and grow, shifting across the grid like evanescent neural impulses or bioluminescent clusters of diatoms. In Stephen Hawking's 2012 film *The Meaning of Life*, the narrator describes Conway's mathematical model as simulating "how a complex thing like the mind might come about from a basic set of rules," revealing the overweening ambitions that characterize contemporary AI: "[T]hese complex properties emerge from simple laws that contain no concepts like movement or reproduction," yet they produce "species," and cells "can even reproduce, just as life does in the real world." [50]

Just as life does? Artists know the blandishments of simulation and representation, the difference between the genius of artifice and the realities of what "life does." Parreno's piece is an intuitive assembly of our experience of "life" through embodied, perspectival engagement. Our consciousness is electrically (cybernetically) enmeshed, yet we don't respond as if this human-generated set of elegant simulations had its own *intelligence*.

---

[50] Narration in Stephen Hawking's *The Meaning of Life* (Smithson Productions, Discovery Channel, 2012).

The artistic use of cybernetic beings also reminds us that consciousness itself is not just "in here." It is streaming in and out, harmonizing those sensory, scintillating signals. Mind happens well outside the limits of the cranium (and its simulacrum, the "motherboard"). In Mary Catherine Bateson's paraphrase of her father Gregory's second-order cybernetics, mind is material "not necessarily defined by a boundary such as an envelope of skin."[51] Parreno pairs the simulations of art with the simulations of mathematics to force the Wiener-like point that any such model is not, by itself, just like life. Models are just that—parts of signaling systems constituting "intelligence" only when their creaturely counterparts engage them in lively meaning making. Contemporary AI has talked itself into a corner by instrumentalizing and particularizing tasks and subroutines, confusing these drills with actual wisdom. The brief cultural history offered here reminds us that views of data as intelligence, digital nets as "neural," or isolated individuals as units of life, were alien even to Conway's brute simulation.

We can stigmatize the stubborn arrogance of current AI as "right cybernetics," the path that led to current automated weapons systems, Uber's ill-disguised hostility to human workers, and the capitalist dreams of Google. Now we must turn back to left cybernetics—theoretical biologists and anthropologists engaged with a trans-species understanding of intelligent systems. Gregory Bateson's observation that corporations merely simulate "aggregates of parts of persons," with profit-maximizing decisions cut off from "wider and wiser parts of the mind," has never been more timely.[52]

The cybernetic epistemology offered here suggests a new approach. The individual mind is immanent, not only in the body but also in pathways outside the body, and there is a larger Mind, of which the individual mind is only a subsystem. This larger Mind, Bateson holds, is comparable to God, and is perhaps what some people mean by "God," but it is still immanent in the total interconnected social system and planetary ecology. This is not the collective delusion of an exterior "God" who speaks from outside human consciousness (this long-seated monotheistic conceit, Bateson suggests, leads to views of nature and environment as also outside the "individual" human, rendering them as "gifts to exploit"). Rather, Bateson's "God" is a placeholder for our evanescent experience of interacting consciousness-in-the-world: larger Mind as a result of inputs and actions that then become inputs for other actions in concert with other entities—webs of symbiotic relationships that form patterns we need urgently to sense and harmonize with.[53]

From Tsai in the 1970s to Hershman Leeson in the 1990s to Parreno in 2014, artists have been critiquing right cybernetics and plying alternative, embodied, environmental experiences of "artificial" intelligence. Their artistic use of cybernetic beings offers the wisdom of symbionts experienced in the kinds of poeisis that can be achieved in this world: rhythms of signals and intuitive actions that produce the movements of life partnered with an electro-mechanical and -magnetic technosphere. Life, in its mysterious negentropic entanglements with matter and Mind.

---

[51] Mary Catherine Bateson, 1999 foreword to Gregory Bateson, *Steps to an Ecology of Mind* (Chicago: University of Chicago Press, 1972): xi.

[52] *Steps to an Ecology of Mind,* p. 452.

[53] Ibid., pp. 467-8.

Over nearly four decades, **Stephen Wolfram** *has been a pioneer in the development and application of computational thinking and responsible for many innovations in science, technology and business.*

His 1982 paper "Cellular Automata as Simple Self-Organizing Systems," written at the age of twenty-three, was the first of numerous significant scientific contributions aimed at understanding the origins of complexity in nature.

It was around this time that Stephen briefly came into my life. I had established The Reality Club, an informal gathering of intellectuals who met in New York City to present their work before peers in other disciplines. (Note: In 1996, The Reality Club went online as Edge.org). *Our first speaker? Stephen Wolfram, a "wunderkind" who had arrived in Princeton at the Institute for Advanced Study. I distinctly recall his focused manner as he sat down on a couch in my living room and spoke uninterrupted for about an hour before the assembled group.*

Since that time, Stephen has become intent making the world's knowledge easily computable and accessible. His program Mathematica *is the definitive system for modern technical computing.* Wolfram|Alpha *computes expert-level answers using AI technology. He considers his* Wolfram Language *to be the first true computational communication language for humans and AIs.*

I caught up with him again four years ago, when we arranged to meet in Cambridge, Massachusetts, for a freewheeling conversation about AI. Stephen walked in, said hello, sat down, and, looking at the video camera set up to record the conversation for Edge, *began to talk and didn't stop for two and a half hours.*

The essay that follows is an edited version of that session, which was a Wolfram master class of sorts and is an appropriate way to end this volume—just as Stephen's Reality Club talk in the '80s was a great way to initiate the ongoing intellectual enterprise whose result is the rich community of thinkers presenting their work to one another and to the public in this book.

ARTIFICIAL INTELLIGENCE AND THE FUTURE OF CIVILIZATION

Stephen Wolfram

***Stephen Wolfram** is a scientist, inventor, and the founder and CEO of Wolfram Research. He is the creator of the symbolic computation program* Mathematica *and its programming language,* Wolfram Language*, as well as the knowledge engine* Wolfram|Alpha. *He is also the author of* A New Kind of Science.

The following is an edited transcript from a live interview with him conducted in December 2015.

I see technology as taking human goals and making them automatically executable by machines. Human goals of the past have entailed moving objects from here to there, using a forklift rather than our own hands. Now the work we can do automatically, with machines, is mental rather than physical. It's obvious that we can automate many of the tasks we humans have long been proud of doing ourselves. What's the future of the human condition in that situation?

People talk about the future of intelligent machines and whether they'll take over and decide what to do for themselves. But the inventing of goals is not something that has a path to automation. Someone or something has to define what a machine's purpose should be—what it's trying to execute. How are goals defined? For a given human, they tend to be defined by personal history, cultural environment, the history of our civilization. Goals are uniquely human. Where the machine is concerned, we can give it a goal when we build it.

What kinds of things have intelligence, or goals, or purpose? Right now, we know one great example, and that's us—our brains, our human intelligence. Human intelligence, I once assumed, is far beyond anything else that exists naturally in the world; it's the result of an elaborate process of evolution and thus stands apart from the rest of existence. But what I've realized, as a result of the science I've done, is that this is not the case.

People might say, for instance, "The weather has a mind of its own." That's an animist statement and seems to have no place in modern scientific thinking. But it's not as silly as it sounds. What does the human brain do? A brain receives certain input, it computes things, it causes certain actions to happen, it generates a certain output. Like the weather. All sorts of systems are, effectively, doing computations—whether it's a brain or, say, a cloud responding to its thermal environment.

We can argue that our brains are doing vastly more sophisticated computations than those in the atmosphere. But it turns out that there's a broad equivalence between the kinds of computations that different kinds of systems do. This renders the question of the human condition somewhat poignant, because it seems we're not as special as we thought. There are all those different systems of nature that are pretty much equivalent, in terms of their computational capabilities.

What makes us different from all those other systems is the particulars of our history, which give us our notions of purpose and goals. That's a long way of saying that when the box on our desk thinks as well as the human brain does, what it still won't have, intrinsically, are goals and purposes. Those are defined by our particulars—our particular biology, our particular psychology, our particular cultural history.

When we consider the future of AI, we need to think about the goals. That's what humans contribute; that's what our civilization contributes. The execution of those goals is what we can increasingly automate. What will the future of humans be in such a world? What will there be for them to do? One of my projects has been to understand the evolution of human purposes over time. Today we've got all kinds of purposes. If you look back a thousand years, people's goals were quite different: How do I get my food? How do I keep myself safe? In the modern Western world, for the most part you don't spend a large fraction of your life thinking about those purposes. From the point of view of a thousand years ago, some of the goals people have today would seem utterly bizarre—for example, like exercising on a treadmill. A thousand years ago that would sound like a crazy thing to do.

What will people be doing in the future? A lot of purposes we have today are generated by scarcity of one kind or another. There are scarce resources in the world. People want to get more of something. Time itself is scarce in our lives. Eventually, those forms of scarcity will disappear. The most dramatic discontinuity will surely be when we achieve effective human immortality. Whether this will be achieved biologically or digitally isn't clear, but inevitably it will be achieved. Many of our current goals are driven in part by our mortality: "I'm only going to live a certain time, so I'd better get this or that done." And what happens when most of our goals are executed automatically? We won't have the kinds of motivations we have today. One question I'd like an answer for is, What do the derivatives of humans in the future end up choosing to do with themselves? One of the potential bad outcomes is that they just play video games all the time.

~ ~ ~

The term "artificial intelligence" is evolving, in its use in technical language. These days, AI is very popular, and people have some idea of what it means. Back when computers were being developed, in the 1940s and 1950s, the typical title of a book or a magazine article about computers was "Giant Electronic Brains." The idea was that just as bulldozers and steam engines and so on automated mechanical work, computers would automate intellectual work. That promise turned out to be harder to fulfill than many people expected. There was, at first, a great deal of optimism; a lot of government money got spent on such efforts in the early 1960s. They basically just didn't work.

There are a lot of amusing science-fiction-ish portrayals of computers in the movies of that time. There's a cute one called *Desk Set,* which is about an IBM-type computer being installed in a broadcasting company and putting everybody out of a job. It's cute because the computer gets asked a bunch of reference-library questions. When my colleagues and I were building Wolfram|Alpha, one of the ideas we had was to get it to answer all of those reference-library questions from *Desk Set*. By 2009, it could answer them all.

In 1943, Warren McCulloch and Walter Pitts came up with a model for how brains conceptually, formally, might work—an artificial neural network. They saw that their brainlike model would do computations in the same way as Turing Machines. From their work, it emerged that we could make brainlike neural networks that would act as general computers. And in fact, the practical work done by the ENIAC folks and John von Neumann and others on computers came directly not from Turing Machines but through this bypath of neural networks.

But simple neural networks didn't do much. Frank Rosenblatt invented a learning device he called the perceptron, which was a one-layer neural network. In the late sixties, Marvin Minsky and Seymour Papert wrote a book titled *Perceptrons*, in which they basically proved that perceptrons couldn't do anything interesting, which is correct. Perceptrons could only make linear distinctions between things. So the idea was more or less dropped. People said, "These guys have written a proof that neural networks can't do anything interesting, therefore no neural networks can do anything interesting, so let's forget about neural networks." That attitude persisted for some time.

Meanwhile, there were a couple of other approaches to AI. One was based on understanding, at a formal level, symbolically, how the world works; and the other was based on doing statistics and probabilistic kinds of things. With regard to symbolic AI, one of the test cases was, Can we teach a computer to do something like integrals? Can we teach a computer to do calculus? There were tasks like machine translation, which people thought would be a good example of what computers could do. The bottom line is that by the early seventies, that approach had crashed.

Then there was a trend toward devices called expert systems, which arose in the late seventies and early eighties. The idea was to have a machine learn the rules that an expert uses and thereby figure out what to do. That petered out. After that, AI became little more than a crazy pursuit.

~ ~ ~

I had been interested in how you make an AI-like machine since I was a kid. I was interested particularly in how you take the knowledge we humans have accumulated in our civilization and automate answering questions on the basis of that knowledge. I thought about how you could do that symbolically, by building a system that could break down questions into symbolic units and answer them. I worked on neural networks at that time and didn't make much progress, so I put it aside for a while.

Back in mid-2002 to 2003, I thought about that question again: What does it take to make a computational knowledge system? The work I'd done by then pretty much showed that my original belief about how to do this was completely wrong. My original belief had been that in order to make a serious computational knowledge system, you first had to build a brainlike device and then feed it knowledge—just as humans learn in standard education. Now I realized that there wasn't a bright line between what is intelligent and what is simply computational.

I had assumed that there was some magic mechanism that made us vastly more capable than anything that was just computational. But that assumption was wrong. This insight is what led to Wolfram|Alpha. What I discovered is that you can take a large collection of the world's knowledge and automatically answer questions on the basis of it, using what are essentially merely computational techniques. It was an alternative way to do engineering—a way that's much more analogous to what biology does in evolution.

In effect, what you normally do when you build a program is build it step-by-step. But you can also explore the computational universe and mine technology from that universe. Typically, the challenge is the same as in physical mining: That is, you find a supply of, let's say, iron, or cobalt, or gadolinium, with some special magnetic properties, and you turn that special capability to a human purpose, to something you want technology to do. In the case of magnetic materials, there are plenty of ways to do that. In terms of programs, it's the same story. There are all kinds of programs out there, even tiny programs that do complicated things. Could we entrain them for some useful human purpose?

And how do you get AIs to execute your goals? One answer is to just talk to them, in the natural language of human utterances. It works pretty well when you're talking to Siri. But when you want to say something longer and more complicated, it doesn't work well. You need a computer language that can represent sophisticated concepts in a way that can be progressively built up and isn't possible in natural language. What my company spent a lot of time doing was building a knowledge-based language that incorporates the knowledge of the world directly into the language. The traditional approach to creating a computer language is to make a language that represents operations that computers intrinsically know how to do: allocating memory, setting values of variables, iterating things, changing program counters, and so on. Fundamentally, you're telling computers to do things in your own terms. My approach was to make a language that panders not to the computers but to the humans, to take whatever a human thinks of and convert it into some form that the computer can understand. Could we encapsulate the knowledge we'd accumulated, both in science and in data collection, into a language we could use to communicate with computers? That's the big achievement of my last thirty years or so—being able to do that.

Back in the 1960s, people would say things like, "When we can do such-and-such, we'll know we have AI. When we can do an integral from a calculus course, we'll know we have AI. When we can have a conversation with a computer and make it seem human. . . ," et cetera. The difficulty was, "Well, gosh, the computer just doesn't know enough about the world." You'd ask the computer what day of the week it was, and it might be able to answer that. You'd ask it who the President was, and it probably couldn't tell you. At that point, you'd know you were talking to a computer and not a person. But now when it comes to these Turing Tests, people who've tried connecting, for example, Wolfram|Alpha to their Turing Test bots find that the bots lose every time. Because all you have to do is start asking the machine sophisticated questions and it will answer them! No human can do that. By the time you've asked it a few disparate questions, there will be no human who knows all those things, yet the system will know them. In that sense, we've already achieved good AI, at that level.

Then there are certain kinds of tasks easy for humans but traditionally very hard for machines. The standard one is visual object identification: What is this object? Humans can recognize it and give some simple description of it, but a computer was just hopeless at that. A couple of years ago, though, we brought out a little image-identification system, and many other companies have done something similar—ours happens to be somewhat better than the rest. You show it an image, and for about ten thousand kinds of things, it will tell you what it is. It's fun to show it an abstract painting and see what it says. But it does a pretty good job.

It works using the same neural-network technology that McCulloch and Pitts imagined in 1943 and lots of us worked on in the early eighties. Back in the 1980s, people successfully did OCR—optical character recognition. They took the twenty-six letters of the alphabet and said, "OK, is that an A? Is that a B? Is that a C?" and so on.

That could be done for twenty-six different possibilities, but it couldn't be done for ten thousand. It was just a matter of scaling up the whole system that makes this possible today. There are maybe five thousand picturable common nouns in English, ten thousand if you include things like special kinds of plants and beetles which people would recognize with some frequency. What we did was train our system on 30 million images of these kinds of things. It's a big, complicated, messy neural network. The details of the network probably don't matter, but it takes about a quadrillion GPU operations to do the training.

Our system is impressive because it pretty much matches what humans can do. It has about the same training data humans have—about the same number of images a human infant would see in the first couple of years of its life. Roughly the same number of operations have to be done in the learning process, using about the same number of neurons in at least the first levels of our visual cortex. The details are different; the way these artificial neurons work has little to do with how the brain's neurons work. But the concept is similar, and there's a certain universality to what's going on. At the mathematical level, it's a composition of a very large number of functions, with certain continuity properties that let you use calculus methods to incrementally train the system. Given those attributes, you can end up with something that does the same job human brains do in physiological recognition.

But does this constitute AI? There are a few basic components. There's physiological recognition, there's voice-to-text, there's language translation—things humans manage to do with varying degrees of difficulty. These are essentially some of the links to how we make machines that are humanlike in what they do. For me, one of the interesting things has been incorporating those capabilities into a precise symbolic language to represent the everyday world. We now have a system that can say, "This is a glass of water." We can go from a picture of a glass of water to the *concept* of a glass of water. Now we have to invent some actual symbolic language to represent those concepts.

I began by trying to represent mathematical, technical kinds of knowledge and went on to other kinds of knowledge. We've done a pretty good job of representing objective knowledge in the world. Now the problem is to represent everyday human discourse in a precise symbolic way—a knowledge-based language intended for communication between humans and machines, so that humans can read it and machines can understand it, too. For instance, you might say "X is greater than 5." That's a predicate. You might also say, "I want a piece of chocolate." That's also a predicate. It has an "I want" in it. We have to find a precise symbolic representation of the desires we express in human natural language.

In the late 1600s, Gottfried Leibniz, John Wilkins, and others were concerned with what they called philosophical languages—that is, complete, universal, symbolic representations of things in the world. You can look at the philosophical language of John Wilkins and see how he divided up what was important in the world at the time. Some aspects of the human condition have been the same since the 1600s. Some are very different. His section on death and various forms of human suffering was huge; in today's ontology, it's a lot smaller. It's interesting to see how a philosophical language of today would differ from a philosophical language of the mid-1600s. It's a measure of our progress. Many such attempts at formalization have happened over the years. In mathematics, for example: Whitehead and Russell's *Principia Mathematica* in 1910 was the biggest showoff effort. There were previous attempts by Gottlob Frege and Giuseppe Peano that were a little more modest in their presentation. Ultimately, they were wrong in what they thought they should formalize: They thought they should formalize some process of mathematical proof, which turns out not to be what most people care about.

With regard to a modern analog of the Turing Test, it's an interesting question. There's still the conversational bot, which is Turing's idea. That one hasn't been solved yet. It will be solved—the only question is, What is the application for which it is solved? For a long time I would ask, "Why should we care?"—because I thought the principal application would be customer service, which wasn't particularly high on my list. But customer service, where you're trying to interface, is just where you need this conversational language.

One big difference between Turing's time and ours is the method of communicating with computers. In his time, you typed something into the machine and it typed back a response. In today's world, it responds with a screen—as for instance, when you want to buy a movie ticket. How is a transaction with a machine different from a transaction with a human? The main answer is that there's a visual display. It asks you something, and you press a button, and you can see the result immediately. For example, in Wolfram|Alpha, when it's used inside Siri, if there's a short answer, Siri will tell you the short answer. But what most people want is the visual display, showing the infographic of this or that. This is a nonhuman form of communication that turns out to be richer than the traditional spoken, or typed, human communication. In most human-to-human communication, we're stuck with pure language, whereas in computer-to-human communication we have this much higher bandwidth channel—of visual communication.

Many of the most powerful applications of the Turing Test fall away now that we have this additional communication channel. For example, here's one we're pursuing right now. It's a bot that communicates about writing programs: You say, "I want to write a program. I want it to do this." The bot will say, "I've written this piece of program. This is what it does. Is this what you want?" Blah-blah-blah. It's a back-and-forth bot. Devising such systems is an interesting problem, because they have to have a model of a human if they're trying to explain something to you. They have to know what the human is confused about.

What has long been difficult for me to understand is, What's the point of a conventional Turing Test? What's the motivation? As a toy, one could make a little chat bot that people could chat with. That will be the next thing. The current round of deep learning—particularly, recurrent neural networks—is making pretty good models of human speech and human writing. We can type in, say, "How are you feeling today?" and it knows most of the time what sort of response to give. But I want to figure out whether I can automate responding to my email. I know the answer is "No." A good Turing Test, for me, will be when a bot can answer most of my email. That's a tough test. It would have to learn those answers from the humans the email is connected to. I might be a little bit ahead of the game, because I've been collecting data on myself for about twenty-five years. I have every piece of email for twenty-five years, every keystroke for twenty. I should be able to train an avatar, an AI, that will do what I can do—perhaps better than I could.

~ ~ ~

People worry about the scenario in which AIs take over. I think something much more amusing, in a sense, will happen first. The AI will know what you intend, and it will be good at figuring out how to get there. I tell my car's GPS I want to go to a particular destination. I don't know where the heck I am, I just follow my GPS. My children like to remind me that once when I had a very early GPS—the kind that told you, "Turn this way, turn that way"—we ended up on one of the piers going out into Boston Harbor.

More to the point is that there will be an AI that knows your history, and knows that when you're ordering dinner online you'll probably want such-and-such, or when you email this person, you should talk to them about such-and-such. More and more, the AIs will suggest to us what we should do, and I suspect most of the time people will just go along with that. It's good advice—better than what you would have figured out for yourself.

As far as the takeover scenario is concerned, you can do terrible things with technology and you can do good things with technology. Some people will try to do terrible things with technology, and some people will try to do good things with technology. One of the things I like about today's technology is the equalization it has produced. I used to be proud that I had a better computer than anybody I knew; now we all have the same kind of computers. We have the same smartphones, and pretty much the same technology can be used by a decent fraction of the planet's 7 billion people. It's not the case that the king's technology is different from everybody else's. That's an important advance.

The great frontier five hundred years ago was literacy. Today, it's doing programming of some kind. Today's programming will be obsolete in a not very long time. For example, people no longer learn assembly language, because computers are better at writing assembly language than humans are, and only a small set of people need to know the details of how language gets compiled into assembly language. A lot of what's being done by armies of programmers today is similarly mundane. There's no good reason for humans to be writing Java code or JavaScript code. We want to automate the programming process so that what's important goes from what the human wants done to getting the machine, as automatically as possible, to do it. This will increase that equalization, which is something I'm interested in. In the past, if you wanted to write a serious piece of code, or program for something important and real, it was a lot of work. You had to know quite a bit about software engineering, you had to invest months of time in it, you had to hire programmers who knew this or you had to learn it yourself. It was a big investment.

That's not true anymore. A one-line piece of code already does something interesting and useful. It allows a vast range of people who couldn't make computers do things for them, make computers do things for them. Something I'd like to see is a lot of kids around the world learn the new capabilities of knowledge-based programming and then produce code that's effectively as sophisticated as what anybody in the top ranks can produce. This is within reach. We're at the point where anybody can learn to do knowledge-based programming, and, more important, learn to think computationally. The actual mechanics of programming are easy now. What's difficult is imagining things in a computational way.

How do you teach computational thinking? In terms of how to do programming, it's an interesting question. Take nanotechnology. How did we achieve nanotechnology? Answer: We took technology as we understand it on a large scale and we made it very small. How to make a CPU chip on the atomic scale? Fundamentally, we use the same architecture as the CPU chip we know and love. That isn't the only approach one can take. Looking at what simple programs can do suggests that you can take even simple impoverished components and with the right compiler you can make them do interesting things. We don't do molecular-scale computing yet, because the ambient technology is such that you'd have to spend a decade building it. But we've got the components that are enough to make a universal computer. You might not know how to program with those components, but by doing searches in the space of possible programs, you'd start to amass building blocks, and you could then create a compiler for them. The surprising thing is that impoverished stuff is capable of doing sophisticated things, and the compilation step is not as gruesome as you might expect.

Just searching the computational universe and trying to find programs—building blocks—that are interesting is a good approach. A more traditional engineering approach—trying by pure thought to figure out how to build a universal computer—is a harder row to hoe. That doesn't mean it can't be done, but my guess is that we'll be able to do some amazing things just by finding the components and searching the possible programs we can make with them. Then it's back to the question about connecting human purposes to what is available from the system.

One question I'm interested in is, What will the world look like when most people can write code? We had a transition, maybe five hundred years ago or so, when only scribes and a small set of the population could read and write natural language. Today, a small fraction of the population can write code. Most of the code they write is for computers only. You don't understand things by reading code. But there will come a time when, as a result of things I've tried to do, the code is at a high enough level that it's a minimal description of what you're trying to do. It will be a piece of code that's understandable to humans but also executable by the machines.

Coding is a form of expression, just as writing in a natural language is a form of expression. To me, some simple pieces of code are poetic—they express ideas in a very clean way. There's an aesthetic aspect, much as there is to expression in a natural language. One feature of code is that it's immediately executable; it's not like writing. When you write something, somebody has to read it, and the brain that's reading it has to absorb the thoughts that came from the person who did the writing. Look at how knowledge has been transmitted in the history of the world. At level zero, one form of knowledge transmission is essentially genetic—that is, there's an organism, and its progeny has the same features that it had. Then there's the kind of knowledge transmission that happens with things like physiological recognition. A newborn creature has some neural network with some random connections in it, and as the creature moves around in the world, it starts recognizing kinds of objects and it learns that knowledge.

Then there's the level that was the big achievement of our species, which is natural language. The ability to represent knowledge abstractly enough that we can communicate it brain to brain, so to speak. Arguably, natural language is our species' most important invention. It's what led, in many respects, to our civilization.

There's yet another level, and probably one day it will have a more interesting name. With knowledge-based programming, we have a way of creating an actual representation of real things in the world, in a precise and symbolic way. Not only is it understandable by brains and communicable to other brains and to computers, it's also immediately executable.

Just as natural language gave us civilization, knowledge-based programming will give us—what? One bad answer is that it will give us the civilization of the AIs. That's what we don't want to happen, because the AIs will do a great job communicating with one another and we'll be left out of it, because there's no intermediate language, no interface with our brains. What will this fourth level of knowledge communication lead to? If you were Caveman Ogg and you were just realizing that language was starting, could you imagine the coming of civilization? What should we be imagining right now?

This relates to the question of what the world would look like if most people could code. Clearly, many trivial things would change: Contracts would be written in code, restaurant recipes might be written in code, and so on. Simple things like that would change. But much more profound things would also change. The rise of literacy gave us bureaucracy, for example, which had already existed but dramatically accelerated, giving us a greater depth of governmental systems, for better or worse. How does the coding world relate to the cultural world?

Take high school education. If we have computational thinking, how does that affect how we study history? How does that affect how we study languages, social studies, and so on? The answer is, it has a great effect. Imagine you're writing an essay. Today, the raw material for a typical high school student's essay is something that's already been written; students usually can't generate new knowledge easily. But in the computational world, that will no longer be true. If the students know something about writing code, they'll access all that digitized historical data and figure out something new. Then they'll write an essay about something they've discovered. The achievement of knowledge-based programming is that it's no longer sterile, because it's got the knowledge of the world knitted into the language you're using to write code.

~ ~ ~

There's computation all over the universe: in a turbulent fluid producing some complicated pattern of flow, in the celestial mechanics of planetary interactions, in brains. But does computation have a purpose? You can ask that about any system. Does the weather have a goal? Does climate have a goal?

Can someone looking at Earth from space tell that there's anything with a purpose there? Is there a civilization there? In the Great Salt Lake, in Utah, there's a straight line. It turns out to be a causeway dividing two areas of the lake with different colors of algae, so it's a very dramatic straight line. There's a road in Australia that's long and straight. There's a railroad in Siberia that's long, and lights go on when a train stops at the stations. So from space you can see straight lines and patterns.

But are these clear enough examples of obvious purpose on Earth as viewed from space? For that matter, how do we recognize extraterrestrials out there? How do we tell if a signal we're getting indicates purpose? Pulsars were discovered in 1967, when we picked up a periodic flutter every second or so. The first question was, Is this a beacon?

Because what else would make a periodic signal? It turned out to be a rotating neutron star.

One criterion to apply to a potentially purposeful phenomenon is whether it's minimal in achieving a purpose. But does that mean that it was built for the purpose? The ball rolls down the hill because of gravitational pull. Or the ball rolls down the hill because it's satisfying the principle of least action. There are typically these two explanations for some action that seems purposeful: the mechanistic explanation and the teleological. Essentially all of our existing technology fails the test of being minimal in achieving its purpose. Most of what we build is steeped in technological history, and it's incredibly non-minimal for achieving its purpose. Look at a CPU chip; there's no way that that's the minimal way to achieve what a CPU chip achieves.

This question of how to identify purposefulness is a hard one. It's an important question, because radio noise from the galaxy is very similar to CDMA transmissions from cell phones. Those transmissions use pseudo-noise sequences, which happen to have certain repeatability properties. But they come across as noise, and they're set up as noise, so as not to interfere with other channels. The issue gets messier. If we were to observe a sequence of primes being generated from a pulsar, we'd ask what generated them. Would it mean that a whole civilization grew up and discovered primes and invented computers and radio transmitters and did this? Or is there just some physical process making primes? There's a little cellular automaton that makes primes. You can see how it works if you take it apart. It has a little thing bouncing inside it, and out comes a sequence of primes. It didn't need the whole history of civilization and biology and so on to get to that point.

I don't think there is abstract "purpose," per se. I don't think there's abstract meaning. Does the universe have a purpose? Then you're doing theology in some way. There is no meaningful sense in which there is an abstract notion of purpose. Purpose is something that comes from history.

One of the things that might be true about our world is that maybe we go through all this history and biology and civilization, and at the end of the day the answer is "42," or something. We went through all those 4 billion years of various kinds of evolution and then we got to "42."

Nothing like that will happen, because of computational irreducibility. There are computational processes that you can go through in which there is no way to shortcut that process. Much of science has been about shortcutting computation done by nature. For example, if we're doing celestial mechanics and want to predict where the planets will be a million years from now, we could follow the equations, step-by-step. But the big achievement in science is that we're able to shortcut that and reduce the computation. We can be smarter than the universe and predict the endpoint without going through all the steps. But even with a smart enough machine and smart enough mathematics, we can't get to the endpoint without going through the steps. Some details are irreducible. We have to irreducibly follow those steps. That's why history means something. If we could get to the endpoint without going through the steps, history would be, in some sense, pointless.

So it's not the case that we're intelligent and everything else in the world is not. There's no enormous abstract difference between us and the clouds or us and the cellular automata. We cannot say that this brainlike neural network is qualitatively different from this cellular-automaton system. The difference is a detailed difference. This brainlike neural network was produced by the long history of civilization, whereas the cellular automaton was created by my computer in the last microsecond.

The problem of abstract AI is similar to the problem of recognizing extraterrestrial intelligence: How do you determine whether or not it has a purpose? This is a question I don't consider answered. We'll say things like, "Well, AI will be intelligent when it can do blah-blah-blah." When it can find primes. When it can produce this and that and the other. But there are many other ways to get to those results. Again, there is no bright line between intelligence and mere computation.

It's another part of the Copernican story: We used to think Earth was the center of the universe. Now we think we're special because we have intelligence and nothing else does. I'm afraid the bad news is that that isn't a distinction.

Here's one of my scenarios. Let's say there comes a time when human consciousness is readily uploadable into digital form, virtualized and so on, and pretty soon we have a box of a trillion souls. There are a trillion souls in the box, all virtualized. In the box, there will be molecular computing going on—maybe derived from biology, maybe not. But the box will be doing all kinds of elaborate stuff. And there's a rock sitting next to the box. Inside a rock, there are always all kinds of elaborate stuff going on, all kinds of subatomic particles doing all kinds of things. What's the difference between the rock and the box of a trillion souls? The answer is that the details of what's happening in the box were derived from the long history of human civilization, including whatever people watched on YouTube the day before. Whereas the rock has its long geological history but not the particular history of our civilization.

Realizing that there isn't a genuine distinction between intelligence and mere computation leads you to imagine that future—the endpoint of our civilization as a box of trillion souls, each of them essentially playing a video game, forever. What is the "purpose" of that?

From: Morris, Paul V [redacted]
Sent: 1/15/2019 4:44:33 PM
To: jeffrey E. [jeevacation@gmail.com]
Subject: Opportunity Zones
Attachments: image001.jpg; image002.jpg; image003.jpg; image004.jpg

Importance: High

Hope all well, are you doing anything around these or clients?

In a former warehouse on a dimly lit street in the South Bronx, developers sipping Puerto Rican moonshine listened as a local official urged them to capture a new U.S. tax break by rebuilding the decaying neighborhood.

In Alabama, a young lawyer quit his job after seeing the same tax break's potential to help one of the nation's poorest states. He now spends his days driving his Hyundai from town to town, slideshow at the ready, hoping to connect investors with communities.

And on a conference call with potential clients, a prominent hedge fund executive pitched investments in a boutique hotel in Oakland, which he described as San Francisco's Brooklyn. The project is eligible for the same tax break, designed to help the poor.

Betting on Opportunity Zones

Sales of development sites are surging in areas eligible for tax breaks

Source: Real Capital Analytics

Fervor about opportunity zones is heating up across the U.S. For a limited time, investors who develop real estate or fund businesses in these areas are able to defer capital gains on profits earned elsewhere and completely eliminate them on new investments in 8,700 low-income census tracts. The goal is to reinvigorate these areas. But the question is whether the 2017 tax law will, as U.S. Treasury Secretary Steven Mnuchin predicts, pump $100 billion into places that need it most, or if investors will play it safe by funding projects in a few zones already on the upswing.

QuicktakeWill 'Opportunity Zones' Help the Rich, the Poor or Both?

There's no lack of optimism among officials in shrinking Rust Belt towns, wind-swept Western landscapes and hurricane-ravaged Puerto Rico, who hope to jump-start local economies. The incentives are so flexible they could be used for everything from affordable housing to solar farms.

Yet on the investor side, much of the attention is fixed on how to turn a profit in already thriving areas. They include neighborhoods surrounding Manhattan, Atlantic beach towns drawing vacation-home developers, bedroom communities near Silicon Valley and anomalies like Portland, Oregon, where the entire downtown was deemed eligible for the breaks.

"The phrase I keep thinking of is 'gold rush,' " said Michael Lortz, an accountant who works with developers in Portland. "There's a lot of money from out of town that's coming here."

Already, a policy debate is raging. Backers are urging people to reserve judgment and say the tax breaks have galvanized cities, businesses and investors to think creatively about boosting parts of the country most in need. Critics say the incentives were poorly calibrated and may amount to a boondoggle far in excess of the official $1.6 billion projected cost.

Americans may have to wait months or years to learn which side is right. That's because the law doesn't require investors to disclose projects, making it difficult to tell which areas are benefiting the most.

But there's plenty of evidence that a boom is brewing. Goldman Sachs Group Inc., which already had an investment team focusing on struggling communities, has disclosed about $150 million in projects in recent months. Purchases of sites inside opportunity zones spiked as the tax law took effect, outpacing growth in other areas, according to Real Capital Analytics, which tracks property sales. Altogether, investors spent 62 percent more on properties eligible for tax breaks in the 12 months through September, compared with those in the same census tracts a year earlier, its data show.

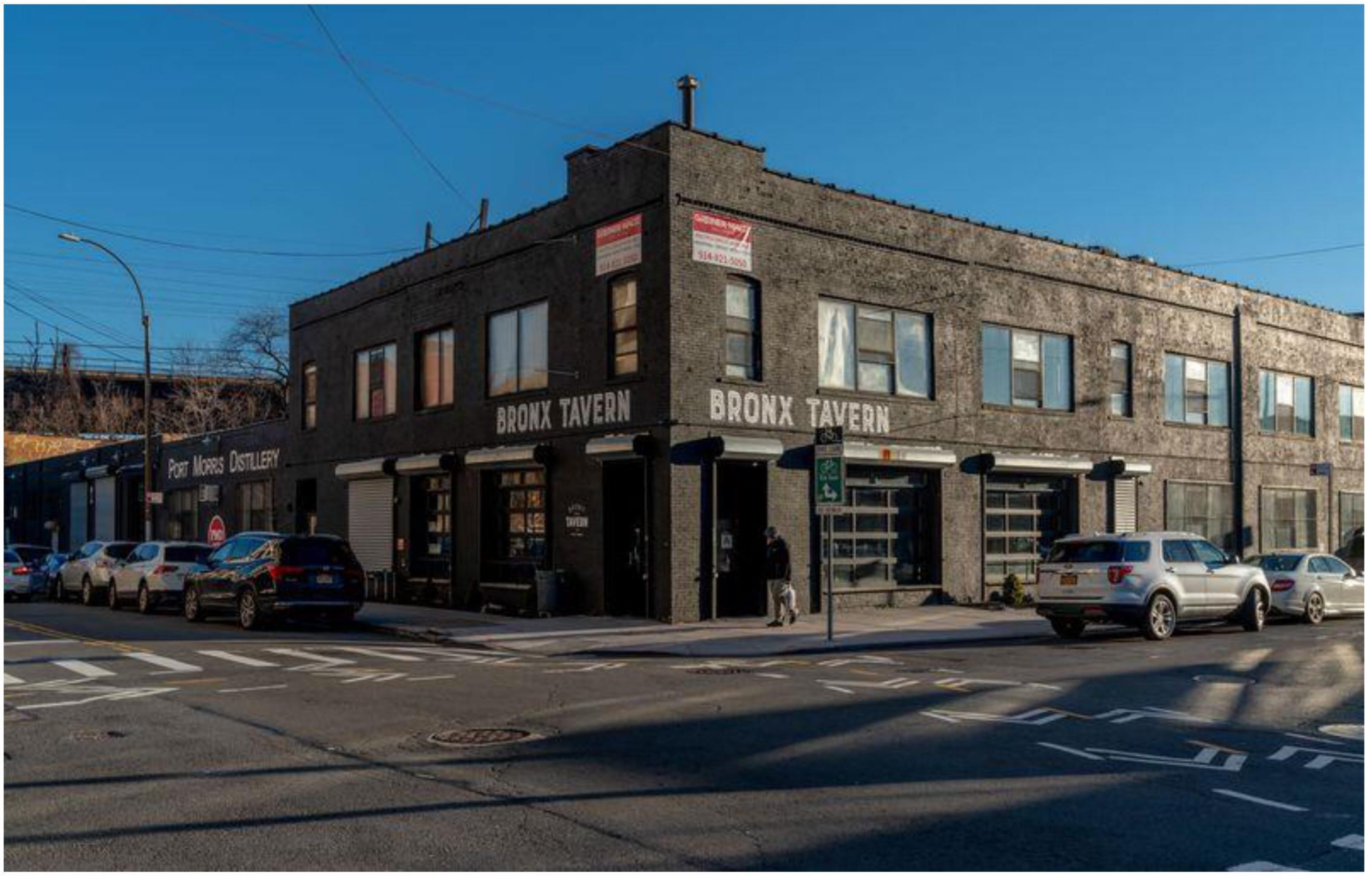
Here are snapshots of what's happening across America:

Bronx Boost

Port Morris Distillery, which makes a high-alcohol rum called pitorro, was the perfect spot for an opportunity zone pitch. It's on a block with a colorful mural, industrial buildings ready for loft conversions and views of the Manhattan skyline. Much of the surrounding South Bronx is now an opportunity zone.

"We have always been the most ignored," Marlene Cintron, the borough's head of economic development, told an audience of developers and lawyers in November. "These opportunity zones are here for you to take advantage of them."

The case for the Bronx, where incomes are among the lowest in New York, is that it's the last borough awaiting revitalization. The tax incentives are designed to unleash it. If developers can buy at current Bronx prices, before seeing a Brooklyn-like rise, the breaks would be massive.



The Bronx Tavern and Port Morris Distillery.
Photographer: David 'Dee' Delgado/Bloomberg

"Huge, huge upside that you're not going to get if you build a strip mall in Topeka," said Terri Adler, managing partner of law firm Duval & Stachenfeld LLP, who also spoke at the event.

Yet the Bronx faces a formidable problem: It's competing with other zones across the city, including waterfronts in Brooklyn and Queens with stronger momentum. The tax break Bronx officials hope will rejuvenate their borough may instead lure more money to what looks like a safer bet across the river.

In November, Amazon.com Inc. selected Long Island City for its next headquarters. Portions of that Queens neighborhood, including a former plastics factory the retailer plans to occupy, are inside opportunity zones, even though they're among the city's fastest-growing areas. In 2017, more apartments were built there than in any other neighborhood in the city.

Driving Alabama

Alex Flachsbart, 30, has a lot of time to talk when he's in his SUV crisscrossing Alabama. A lawyer who specialized in economic development grants and tax breaks, he quit his job last year to start Opportunity Alabama, aiming to connect capital to worthy projects. For the past several months, he's been educating people about opportunity zones, speaking to local officials, businesses and investors.

"I have done the OZ PowerPoint God knows how many times," Flachsbart, who grew up in the Bay Area, said from behind the wheel one day in December.

OPPORTUNITY ZONE TAX BENEFITS
EXIT
THE CITY OF
Athens
ALABAMA

Alex Flachsbart gives a presentation on opportunity zones in Athens, Alabama.
Photographer: Nicole Craine/Bloomberg

When he read about opportunity zones in a 2017 draft of the Tax Cuts and Jobs Act, his mind reeled. Here was an uncapped subsidy far more flexible than anything he'd used before. It could draw investment for an array of projects. He imagined funding startups in Huntsville, where NASA's presence has lured a deep bench of talented engineers, and the renovation of an old civic complex in Mobile.

Flachsbart's nonprofit—which has board members from the state's largest utility and its biggest bank, Regions Financial Corp.—is now in talks for 10 potential projects that need more than $100 million in equity investment, he said. None have been funded yet, but he's certain some will be. Meanwhile, he keeps driving.

Baltimore Billionaire

One of the arguments over opportunity zones is whether the U.S. is handing wealthy investors and companies big breaks on projects they would've done anyway. One example: Hedge fund executive and former White House spokesman Anthony Scaramucci plans to build a "swank, boutique hotel" in Oakland. The paperwork for the permit was filed months before the neighborhood was designated an opportunity zone.

But that project pales in comparison to what's happening in Baltimore. More than a year before President Donald Trump signed the law, real estate developer Steven Siegel helped negotiate one of the largest public financing deals of its kind for a client, a company owned by billionaire Kevin Plank, founder of athletic-wear maker Under Armour Inc.

Port Covington area of Baltimore.
Photographer: Sarah L. Voisin/The Washington Post via Getty Images

In 2016, Baltimore's city council approved a $660 million financing package for a 235-acre mixed-use development, including new offices for Under Armour, along the city's waterfront. The area was already designated as an enterprise zone and a brownfield site, connoting additional lucrative tax breaks, and the project attracted a $233 million investment from Goldman Sachs's urban investment group.

Then came the opportunity zone designation.

The tax break is only supposed to apply to real estate purchased after the law took effect. But lawyers across the country quickly began working around that to get the benefits for projects planned before the law was passed. Many tax experts have recommended sales to new entities. So long as the seller owns no more than 20 percent of the buyer, the transaction counts as arm's-length and qualifies.

Siegel said his firm, Weller Development, has found enough new investors to comply with the arm's-length requirement. The company has seen so much demand, he said, that he's looking to replicate the project elsewhere.

"We've been fielding a lot of inbound interest," Siegel said, declining to name cities that have approached him. "That stimulated us to take this show on the road."

Boulder Balks

In mid-December, during a marathon city council meeting that stretched past midnight, Boulder became perhaps the first jurisdiction in the country to reject its own opportunity zone. Officials in the Colorado town imposed an 18-month moratorium on almost all development in its only census tract earmarked for the incentives.

The move highlights how local officials have the power to respond to criticisms of the law—in this case, that investors may rush to build projects the community doesn't want. Boulder has long been a favorite spot for growing companies because people want to live there, thanks to its college-town vibe and quick access to nature.

"People think we have too much" development, said Bob Yates, a council member. "We're not Detroit."

Sai
Eyebrow Designer

An eyebrow salon is surrounded by empty stores at Diagonal Plaza in Boulder.
Photographer: Rachel Woolf/Bloomberg

Though Yates, a former telecom executive, opposed the moratorium and worries it will make things difficult for businesses in the zone, he understands why people in Boulder, a liberal enclave, were skeptical of an idea enacted by a Republican Congress. He just wishes his city would take a more measured approach.

The zone includes an aging mall called Diagonal Plaza. The tax incentives could have spurred investors to redevelop it to include affordable housing, which is sorely needed, he said.

"Boulder's teachers can't live in Boulder," he said. "It's not a healthy thing to have socioeconomic divide where lower-income people have to live outside of town to serve higher-income people in town."

Paul V. Morris
Managing Director – The Morris Group
Private Banking and Investment Group
Bank of America Merrill Lynch
One Bryant Park, 28th Floor
[redacted]

This message, and any attachments, is for the intended recipient(s) only, may contain information that is privileged, confidential and/or proprietary and subject to important terms and conditions available at http://www.bankofamerica.com/emaildisclaimer. If you are not the intended recipient, please delete this message.

SUPREME COURT OF THE STATE OF NEW YORK
APPELLATE DIVISION – FIRST DEPARTMENT

THE PEOPLE OF THE STATE OF NEW YORK, Respondent, - against - JEFFREY E. EPSTEIN, Defendant-Appellant.	App. Div. No. 6081 On Appeal from New York Supreme Court, New York County, Index No. 30129/10 (Pickholz, J.)

NOTICE OF MOTION BY NYP HOLDINGS, INC. TO UNSEAL APPELLATE BRIEFS

PLEASE TAKE NOTICE that, upon the accompanying affirmation of John M. Browning sworn to on January 11, 2019, together with the exhibits annexed thereto, the accompanying memorandum of law and all pleadings and proceedings heretofore had herein, non-party movant NYP Holdings, Inc., by its undersigned counsel, will move this Court at a term thereof to be held at the Appellate Division Courthouse located at 27 Madison Avenue, New York, New York 10010, on January 21, 2019, at 10:00 a.m., or as soon thereafter as counsel may be heard, for an order unsealing the briefs in the above-captioned appeal.

PLEASE TAKE FURTHER NOTICE that answering papers, if any, shall be served upon the undersigned on or before the last date permitted by CPLR 2214.

Dated: New York, New York
January 11, 2019

Respectfully submitted,

DAVIS WRIGHT TREMAINE LLP

By: ______________________
Robert D. Balin
John M. Browning

1251 Avenue of the Americas, 21st Floor
New York, NY 10020-1104
Telephone: (212) 489-8230
Facsimile: (212) 489-8340
Email: robbalin@dwt.com
johnbrowning@dwt.com

Attorneys for Non-Party Movant NYP Holdings, Inc.

SUPREME COURT, APPELLATE DIVISION
FIRST DEPARTMENT

THE PEOPLE OF THE STATE OF NEW YORK,

Respondent,

- against -

JEFFREY E. EPSTEIN,

Defendant-Appellant.

App. Div. No. 6081

On Appeal from New York
Supreme Court, New York
County, Index No. 30129/10
(Pickholz, J.)

AFFIDAVIT OF SERVICE

STATE OF NEW YORK)
) ss.:
COUNTY OF NEW YORK)

JOHN BROWNING, being duly sworn, states, I am not a party to the action, am over 18 years of age and reside in Kings County, New York. On January 14, 2018, I served a true and correct copy of the annexed by mail in the custody of Federal Express directed to the following:

Martin Weinberg
20 Park Place, Suite 1000
Boston, MA 02116

Karen Friedman Agnifilo
Assistant District Attorney – New York County
One Hogan Place
New York, New York 10013

David Aronberg
State Attorney
15th Judicial Circuit – Palm Beach County Florida
401 North Dixie Highway
West Palm Beach, FL 33401

Dexter Lee
Assistant U.S. Attorney
Southern District of Florida
99 N.E. 4th Street
Miami, FL 33132

Dated: New York, New York

January 11, 2019

Sworn to before me this
11th day of January, 2019

Loretta E Perry
Notary Public

LORETTA E. PERRY
Notary Public, State of New York
No. 24-4931617
Qualified in Kings County
Commission Expires August 1, 2022

By: John Brown
John M. Browning

SUPREME COURT OF THE STATE OF NEW YORK
APPELLATE DIVISION – FIRST DEPARTMENT

THE PEOPLE OF THE STATE OF NEW YORK,

Respondent,

- against -

JEFFREY E. EPSTEIN,

Defendant-Appellant.

App. Div. No. 6081

AFFIRMATION OF JOHN M. BROWNING IN SUPPORT OF MOTION TO UNSEAL

On Appeal from New York Supreme Court,
New York County, Index No. 30129/10
(Pickholz, J.)

John M. Browning, an attorney duly admitted to practice before the Courts of the State of New York, affirms the following under penalties of perjury:

1. I am an associate of Davis Wright Tremaine LLP, attorneys for non-party movant NYP Holdings, Inc., publisher of the New York Post (the "Post") and I submit this affirmation in support of the Post's motion to unseal the briefs filed by the parties in the above-captioned appeal.

2. The grounds for unsealing the appeal briefs in this action are set forth in the accompanying memorandum of law. I submit this affirmation to annex relevant documents and to state facts that are relevant to this motion, of which I have personal knowledge.

3. Annexed hereto as Exhibit A is a true and correct copy of an article written by Julie K. Brown and published by the *Miami Herald* on November 28, 2018, entitled "Cops worked to put serial sex abuser in prison. Prosecutors worked to cut him a break."

4. Annexed hereto as Exhibit B is a true and correct copy of an article written by Rebecca Rosenberg and Danika Fears, which was published by the Post on January 7, 2015, entitled "DA's office 'went easy' on sex offender Epstein."

5. Annexed hereto as Exhibit C is a true and correct copy of an article written by Rebecca Rosenberg, Larry Celona, Susan Edelman and Isabel Vincent, which was published by the Post on December 1, 2018, entitled "Manhattan DA sided with pedophile billionaire after botching investigation."

6. On or about December 4, 2018, Post reporter Susan Edelman contacted Danny Frost, Director of Communications for Manhattan District Attorney Cyrus Vance Jr., and requested copies of the briefs filed by the District Attorney's office in the above-captioned appeal. Ms. Edelman stated that names of victims could be redacted before the briefs were disclosed.

7. Mr. Frost responded that he could not provide Ms. Edelman with copies of the briefs because they were filed under seal pursuant to N.Y. Civil Rights Law 50-b and that the District Attorney's office could only release the briefs, even with victims' names redacted, if this Court ordered the briefs to be unsealed.

8. Mr. Frost further indicated by email that "[i]f the Post petitions the court, and the court asks the People for our position, we will ***not*** oppose the petition for a redacted brief" (emphasis in original). A true and correct copy of that email correspondence is annexed hereto as Exhibit D.

9. On or about December 18, 2018, I contacted Jay Lefkowitz, who represented appellant Jeffrey Epstein in the above-captioned appeal. Mr. Lefkowitz told me that he no longer represents Mr. Epstein and referred me to Martin Weinberg, who currently acts as counsel for Mr. Epstein.

10. On or about December 20, 2018, I spoke with Mr. Weinberg and explained the nature of the Post's motion to unseal the redacted briefs and the position taken by the Manhattan

District Attorney. Mr. Weinberg told me that he was unable to take a position on the Post's motion without first reviewing it and reserved the right to file an opposition, if necessary.

11. On December 21, 2018, the Post filed a motion requesting an order unsealing the appellate briefs and directing the Manhattan District Attorney's Office to produce copies with the names of victims redacted (the "December 21 Motion").

12. On or about December 28, 2018, the Manhattan District Attorney's Office filed an affirmation in response to the Post's motion, signed by Assistant District Attorney Karen Friedman Agnifilo and dated December 28, 2018 (the "DA Affirmation"). A true and correct copy of the Affirmation is annexed hereto as Exhibit E.

13. On January 2, 2019, I spoke with Ms. Friedman Agnifilo to discuss the DA Affirmation. Ms. Friedman Agnifilo told me that there had been a miscommunication by Mr. Frost and that the District Attorney's Office did oppose the December 21 Motion. Ms. Friedman Agnifilo also told me that, in the view of the District Attorney's Office, the Post should give notice of their motion to unseal the appeal briefs to the prosecutors in Florida that handled the prosecution that led to Mr. Epstein's conviction for solicitation of prostitution from a minor. The Post disagrees with that position but nevertheless decided to moot any procedural issues related to requirements to notify the Florida prosecutors. Ms. Friedman Agnifilo subsequently informed me that the agency in Florida responsible for Mr. Epstein's prosecution was the Palm Beach County State Attorney's Office.

14. On January 3, 2019, I received an email from Mr. Weinberg stating that "after consideration of your request for the unsealing of the appellate briefs with redactions of certain identities, we take no position on behalf of Mr. Epstein." A true and correct copy of that email is annexed hereto as Exhibit F.

15. On January 4, 2019, the Post filed a letter with this Court voluntarily withdrawing the December 21 Motion without prejudice to refile in order to moot the procedural issues raised by the Manhattan District Attorney's Office with respect to furnishing notice to prosecutors in Florida. A true and correct copy of that letter is annexed as Exhibit G.

16. On or about January 10, 2019, I contacted the Palm Beach County State Attorney's Office to inform it of the Post's intention to file a motion to unseal the appellate briefs and to inquire about where to send a copy of the motion. A member of the Office of State Attorney David Aronberg instructed me to send a digital copy of the motion to e-postconviction@sa15.org and to send a paper copy to the State Attorney's main offices.

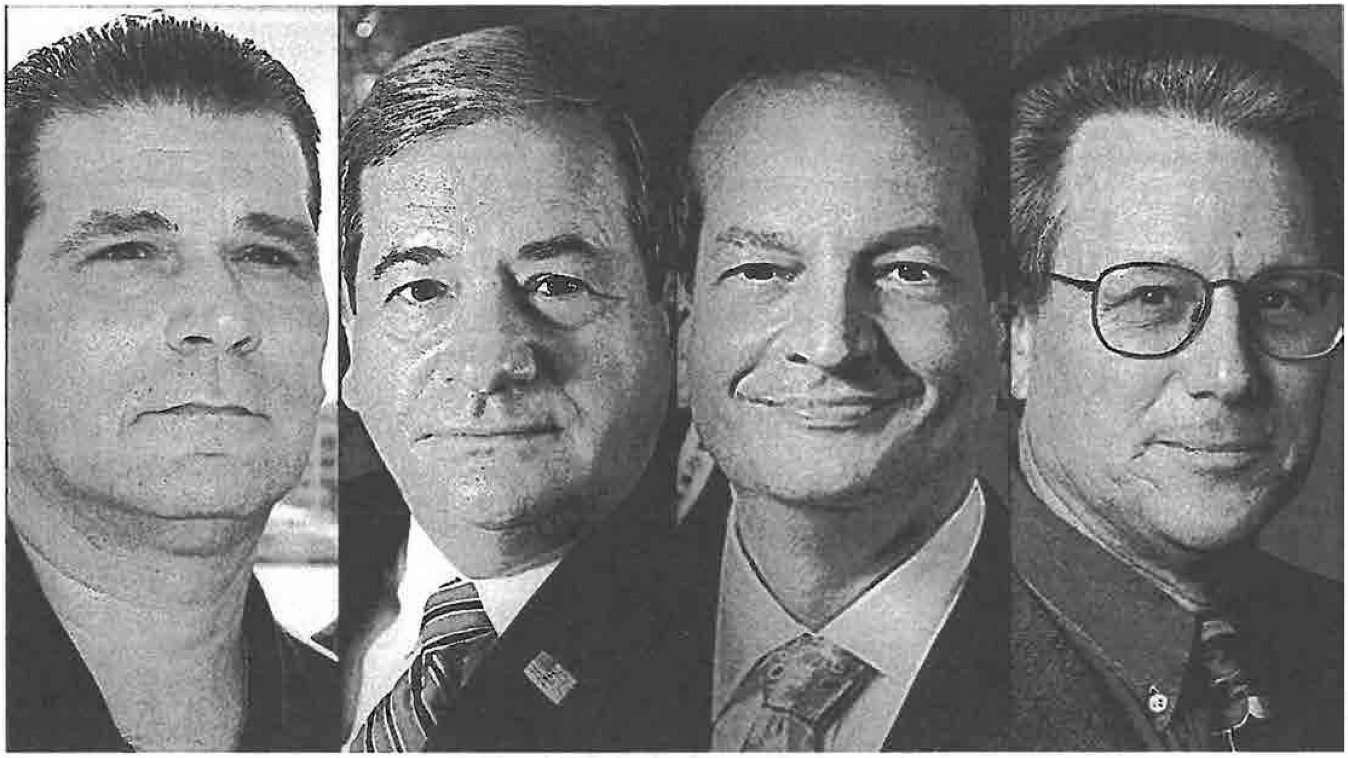
17. On or about January 10 and January 11, 2019, I spoke with Sarah J. Schall, Special Counsel to the U.S. Attorney in the Southern District of Florida, to inform her Office of the Post's intention to file a motion to unseal the appellate briefs and to inquire about where to send a copy of the motion. Ms. Schall instructed me to send a digital copy of the motion papers to her email address and to forward hard copies to Dexter Lee.

Dated: New York, New York
January 11, 2019

John Browning

JOHN M. BROWNING

EXHIBIT A



Cops worked to put serial sex abuser in prison. Prosecutors worked to cut him a break

BY JULIE K. BROWN

NOV. 28, 2018

PERVERSION of JUSTICE

A decade before #MeToo, a multimillionaire sex offender from Florida got the ultimate break.

Palm Beach, Florida

November 2004

Jane Doe

Michelle Licata climbed a narrow, winding staircase, past walls covered with photographs of naked girls. At the top of the stairwell was a vast master bed and bath, with cream-colored shag carpeting and a hot pink and mint green sofa.

The room was dimly lit and very cold.

There was a vanity, a massage table and a timer.

A silver-haired man wearing nothing but a white towel came into the room. He lay facedown on a massage table, and while talking on a phone, directed Licata to rub his back, legs and feet.

Micelle Licata, is one of the over 100 middle school and high school-aged girls that Palm Beach billionaire, Jeffrey Epstein, is accused of sexually assaulting. Licata, now an adult living near Nashville, recalls her experience with

Epstein at his Palm Beach mansion while she was a sophomore at Royal Palm Beach High School.
EMILY MICHOT *EMICHOT@MIAMIHERALD.COM*

After he hung up, the man turned over and dropped his towel, exposing himself. He told Licata to get comfortable and then, in a firm voice, told her to take off her clothes.

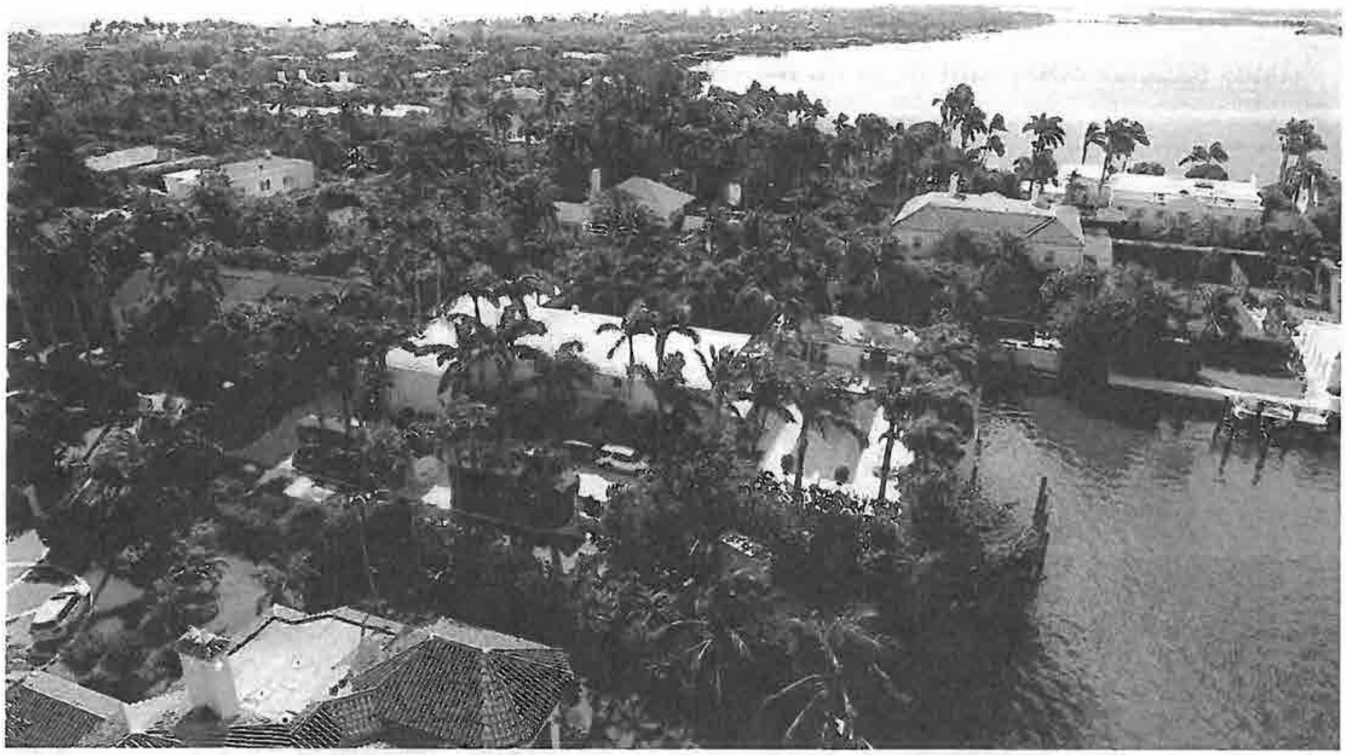
At 16, Licata had never before been fully naked in front of anyone. Shaking and panicked, she mechanically pulled off her jeans and stripped down to her underwear. He set the timer for 30 minutes and then reached over and unsnapped her bra. He then began touching her with one hand and masturbating himself with the other.

"I kept looking at the timer because I didn't want to have this mental image of what he was doing,'' she remembered of the massage. "He kept trying to put his fingers inside me and told me to pinch his nipples. He was mostly saying 'just do that, harder, harder and do this. ...' "

After he ejaculated, he stood up and walked to the shower, dismissing her as if she had been in history class.

It wasn't long before a lot of Licata's fellow students at Royal Palm Beach High School had heard about "a creepy old guy" named Jeffrey who lived in a pink waterfront mansion and was paying girls $200 to $300 to give him massages that quickly turned sexual.

Eventually, the Palm Beach police, and then the FBI, came knocking on Licata's door. In the police report, Licata was referred to as a Jane Doe in order to protect her identity as a minor.



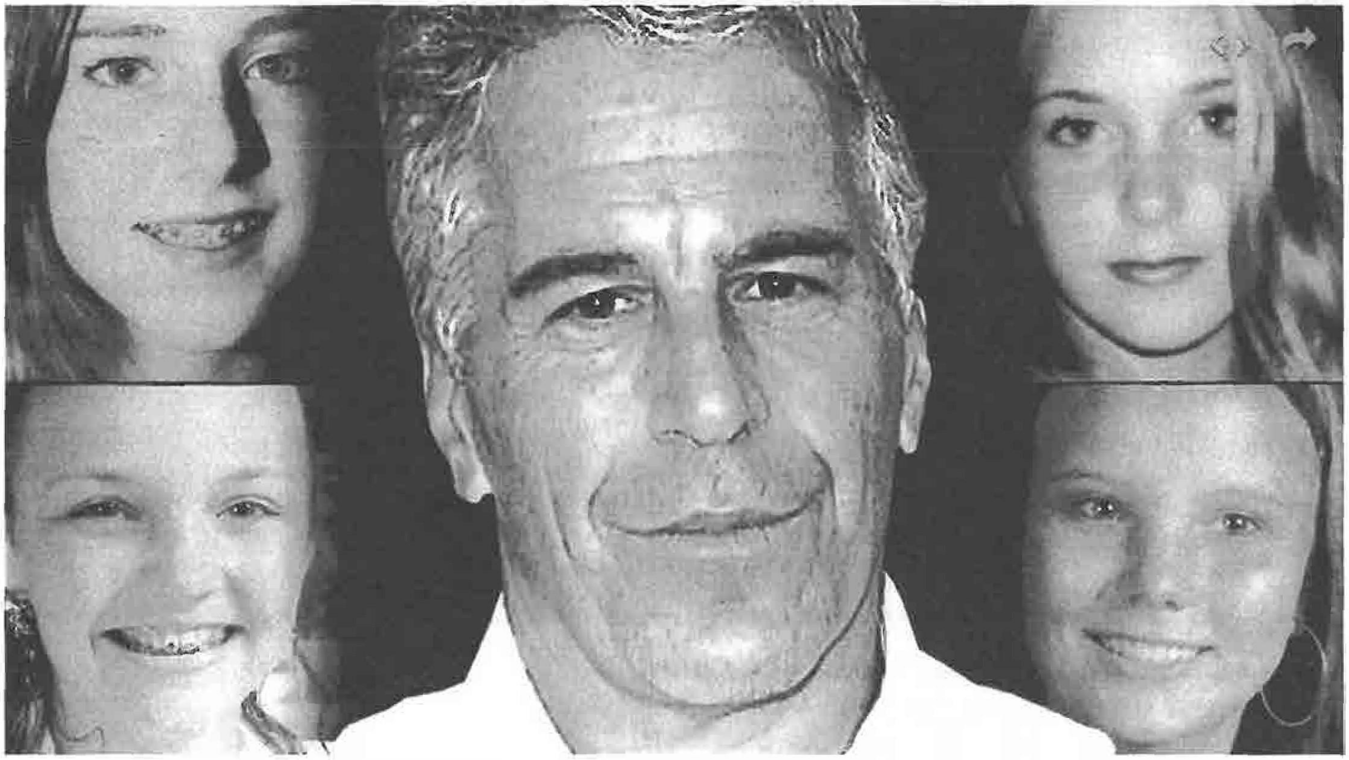
Palm Beach home of registered sex offender, Jeffrey Epstein.
PEDRO PORTAL *PPORTAL@MIAMIHERALD.COM*

There would be many Jane Does to follow: Jane Doe No. 3, Jane Doe No. 4, Jane Does 5, 6, 7, 8 — and as the years went by — Jane Does 102 and 103.

Long before #MeToo became the catalyst for a women's movement about sexual assault — and a decade before the fall of Harvey Weinstein, Bill Cosby and U.S. Olympic gymnastics doctor Larry Nassar — there was Jeffrey Edward Epstein.

Epstein, a multimillionaire hedge fund manager whose friends included a constellation of entertainers, politicians, business titans and royalty, for years lured teenage girls to his Palm Beach mansion as part of a cult-like sex pyramid scheme, police in the town of Palm Beach found.

The girls arrived, sometimes by taxi, for trysts at all hours of the day and night. Few were told much more than that they would be paid to give an old man a massage — and that he might ask them to strip down to their underwear or get naked. But what began as a massage often led to masturbation, oral sex, intercourse and other sex acts, police and court records show. The alleged abuse dates back to 2001 and went on for years.



Palm Beach multimillionaire Jeffrey Epstein is a free man, despite sexually abusing dozens of underage girls according to police and prosecutors. His victims have never had a voice, until now.

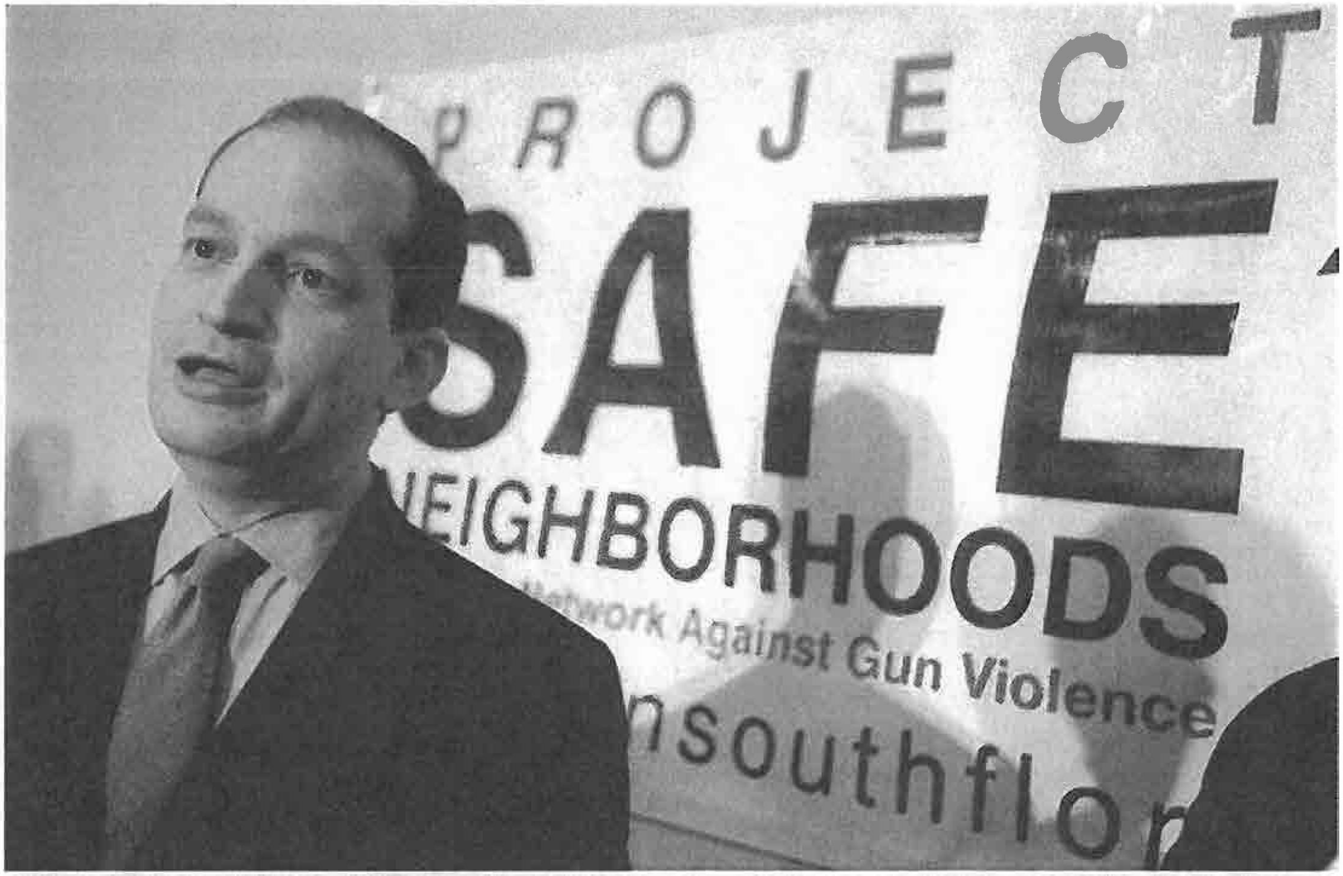
BY EMILY MICHOT ✉ | JULIE K. BROWN ✉

In 2007, despite ample physical evidence and multiple witnesses corroborating the girls' stories, federal prosecutors and Epstein's lawyers quietly put together a remarkable deal for Epstein, then 54. He agreed to plead guilty to two felony prostitution charges in state court, and in exchange, he and his accomplices received immunity from federal sex-trafficking charges that could have sent him to prison for life.

He served 13 months in a private wing of the Palm Beach County stockade. His alleged co-conspirators, who helped schedule his sex sessions, were never prosecuted.

The deal, called a federal non-prosecution agreement, was sealed so that no one — not even his victims — could know the full scope of Epstein's crimes and who else was involved. The U.S. attorney in Miami, Alexander Acosta, was personally involved in the negotiations, records, letters and emails show.

Acosta is now a member of President Donald Trump's Cabinet. As U.S. secretary of labor, he has oversight over international child labor laws and human trafficking and had recently been mentioned as a possible successor to former U.S. Attorney General Jeff Sessions, who resigned under pressure in early November. It was reported on Thursday, a day after this story posted online, that he was no longer in the running.



Alex Acosta, Federal Attorney, speaks during a press conference celebrated during the opening sesion of South Florida Anti-Gang Summit at Miami Hilton Hotel on Sept. 29, 2008 in Miami, Fl. (Cristobal Herrera/Sun Sentinel/TNS)
CRISTOBAL HERRERA *TNS*

The Miami Herald analyzed thousands of pages of court records and lawsuits, witness depositions and newly released FBI documents, and also identified more than 80 women who say they were victimized. They are scattered around the country and abroad. Until now, those victims — today in their late 20s and early 30s — have never spoken publicly about how they felt shamed, silenced and betrayed by the very people in the criminal justice system who were supposed to hold Epstein accountable.

"How come people who don't have money get sent to jail — and can't even make bail — and they have to do their time and sit there and think about what they did wrong? He had no repercussions and doesn't even believe he did anything wrong," said Licata, now 30.

Micelle Licata, is one of the over 100 middle school and high school-aged girls that Palm Beach billionaire, Jeffrey Epstein, is accused of sexually assaulting. Licata, now an adult living near Nashville, recalls her experience with Epstein at his Palm Beach mansion while she was a sophomore at Royal Palm Beach High School.
EMILY MICHOT *EMICHOT@MIAMIHERALD.COM*

Licata is among 36 women who were officially identified by the FBI and the U.S. Attorney's Office as victims of Epstein, now 65. But after the FBI case was closed in 2008, witnesses and alleged victims testified in civil court that there were hundreds of girls who were brought to Epstein's homes, including girls from Europe, Latin America and former Soviet Republic countries.

But Acosta and Epstein's armada of attorneys — Harvard professor Alan Dershowitz, Jay Lefkowitz, Gerald Lefcourt, Jack Goldberger, Roy Black, Guy Lewis and former Whitewater special prosecutor Kenneth Starr — reached a consensus: Epstein would never serve time in a federal or state prison.

READ NEXT

LOCAL

Sex abuser Jeffrey Epstein was surrounded by powerful people. Here's a sampling

NOVEMBER 28, 2018 8:00 AM

POLICE UNDER PRESSURE

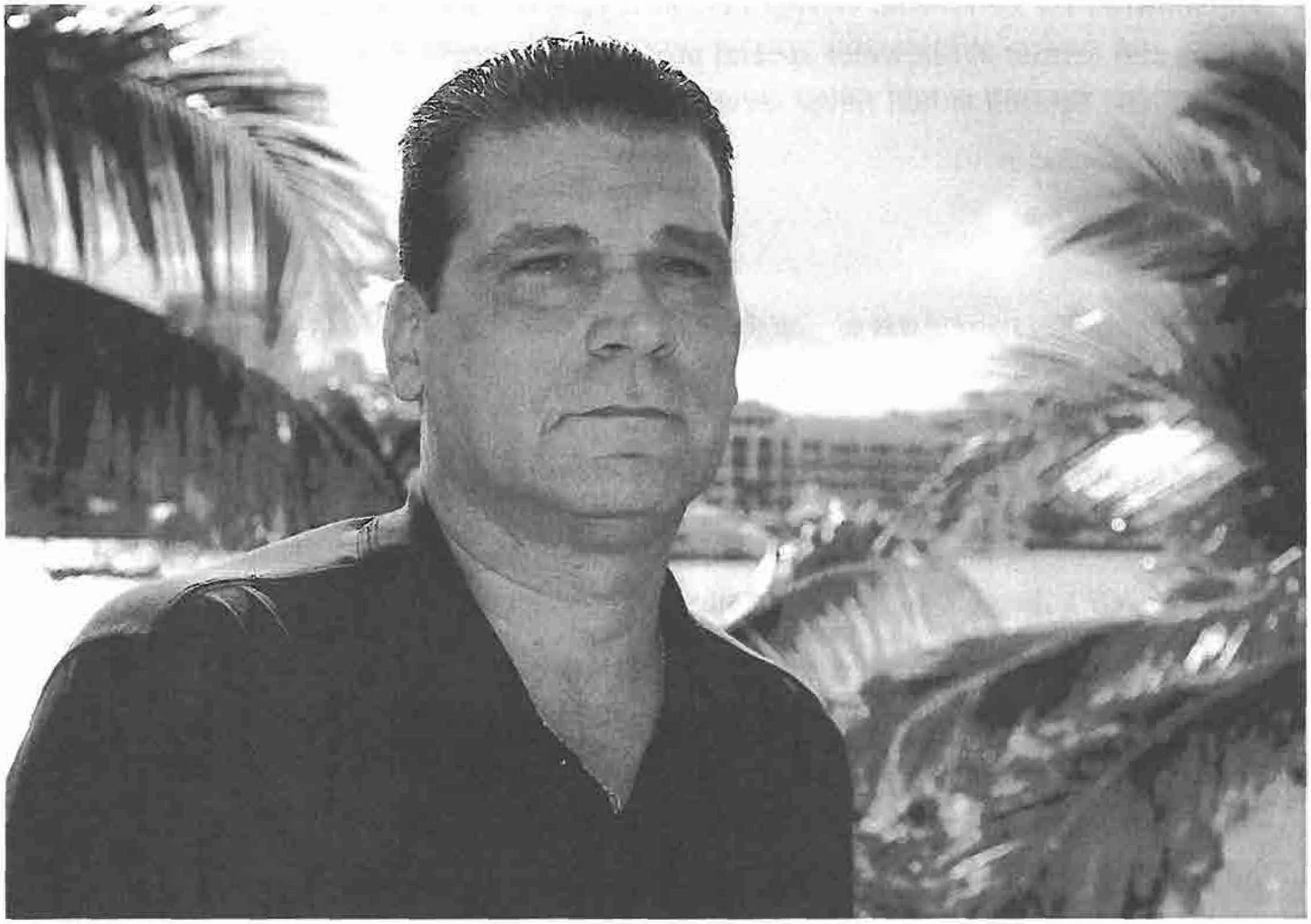
There were really just two people willing to risk their careers to go after Epstein: Palm Beach Police Chief Michael Reiter and Detective Joseph Recarey.

For Reiter, business tycoon Jeffrey Epstein wasn't any more formidable than any of the other 8,000 or so wealthy and powerful people living on the island. Police had handled sensational cases involving wealthy residents before — from the murders of heiresses to the rape case involving William Kennedy Smith, of the Kennedy family.

The easternmost town in Florida, Palm Beach is a 10.4-square-mile barrier island between the Intracoastal Waterway and the Atlantic Ocean populated by some of the richest people in the country. President Trump has his "winter White House" in Palm Beach, and the town makes news as much for its glitz as it does for its unusual efforts to preserve its well-mannered image, like banning shirtless joggers.

But it was a little surprising, even to Reiter, to learn that one of its residents had a revolving door of middle and high school girls coming to his gated compound throughout the day and night.

In their first on-the-record media interviews about the case, Reiter and Recarey revealed new details about the investigation, and how they were, in their view, pressured by then-Palm Beach State Attorney Barry Krischer to downgrade the case to a misdemeanor or drop it altogether.



Former Palm Beach County Police Detective Joe Recarey was the lead detective on the solicitation-of-minors case against billionaire Jeffrey Epstein.
EMILY MICHOT *EMICHOT@MIAMIHERALD.COM*

Between March of 2005 — when the case was opened — and seven months later, when police executed a search warrant at Epstein's home, Recarey had identified 21 possible victims, according to a copy of the unredacted police report obtained by the Herald. By the time police felt they had enough evidence to arrest Epstein on sex charges, they had identified about 35 possible underage victims and were tracking down at least a dozen more, the police report said.

"I was surprised at how quickly it snowballed. I thought at some point there would be a last interview, but the next victim would supply me with three or four more names and the next one had three or four names and it just kept getting bigger and bigger," Recarey said.

By then, word had gotten back to Epstein from some of the girls that they had been questioned by police. Epstein hired famed lawyer Alan Dershowitz.

"Alan Dershowitz flew down and met privately with Krischer," Recarey said. "And the shenanigans that happened, I don't think I've ever seen or heard of before."

Police reports show that Epstein's private investigators attempted to conduct interviews while posing as cops; that they picked through Reiter's trash in search of dirt to discredit him; and that the private investigators were accused of following the girls and their families. In one case, the father of one girl claimed he had been run off the road by a private investigator, police and court reports show.

Support investigative journalism

The Miami Herald obtained thousands of FBI and court records, lawsuits, and witness depositions, and went to federal court in New York to access sealed documents in the reporting of "Perversion of Justice." The Herald also tracked down more than 60 women who said they were victims, some of whom had never spoken of the abuse before.

Your digital subscription, starting at $0.99 for the first month, supports investigative journalism like this.

CLICK TO SUBSCRIBE

Several of the girls said they felt intimidated and frightened by Epstein and Sarah Kellen, the millionaire's assistant and alleged scheduler of massages, who warned them not to talk to police, according to the police report.

Dershowitz, in an interview with the Herald, said he had nothing to do with gathering background on the girls — or in directing anyone to follow the police, or the girls and their families.

"I'm not an investigator. My only job was to negotiate and try the case when it comes to trial," he said.

He nevertheless convinced Krischer that the girls would not be credible on the witness stand, according to Reiter and Recarey.

The defense team's investigators compiled dossiers on the victims in an effort to show that Epstein's accusers had troubled pasts.

Dershowitz met with Krischer and Recarey, sharing with them the results of an investigation into one of the girls, described by Dershowitz as "an accomplished drama student" who hurled profanities at his investigator at "a furious pace."

10/27/99 - OPINION - Barry Krischer; P.B.Co. State Attorney. (AM)

"Our investigation had discovered at least one of her websites and I am enclosing some examples ... the site goes on to detail, including photos, her apparent fascination with marijuana, " Dershowitz wrote in an undated letter to Recarey. He also disputed the claim that one of the defense team's private investigators had misrepresented himself as a police officer.

Recarey stood his ground.

"His attorneys showed us a MySpace page where one of the girls was holding a beer in her hand, and they said, 'oh look, she is underage drinking,' " Recarey recalled. "Well, tell me what teenager doesn't? Does that mean she isn't a victim because she drank a beer? Basically, what you're telling me is the only victim of a sexual battery could be a nun."

Krischer and the lead state prosecutor on the case, Assistant State Attorney Lanna Belohlavek, began to dodge Recarey and Reiter's phone calls and emails, and they dragged their feet on approving subpoenas, Reiter and Recarey said.

"Early on, it became clear that things had changed, from Krischer saying, 'we'll put this guy away for life,' to 'these are all the reasons why we aren't going to prosecute this,' " Reiter said.

Krischer, who is now retired and in private practice, did not respond to multiple requests from the Herald for comment. Belohlavek also did not respond to an email sent to her office.

"It became apparent to me that some of our evidence was being leaked to Epstein's lawyers, who began to question everything that we had in our probable cause affidavit," Reiter said.

The day of the search on Oct. 20, 2005, they found that most of Epstein's computer hard drives, surveillance cameras and videos had been removed from the house, leaving loose, dangling wires, according to the police report.

But the girls' description of the house squared with what detectives found, right down to the hot pink couch and the dresser drawer of sex toys in Epstein's bathroom.

Reiter said his own trash was disappearing from his house, as his life was put under Epstein's microscope. Private investigators hired by Epstein's lawyers even tracked down Reiter's grade school teachers, the former chief said. Questions were raised about donations that Epstein had made to the police department, even though Reiter had returned one of the donations shortly after the investigation began.

Recarey, meanwhile, said he began to take different routes to and from work, and even switched vehicles because he knew he was being tailed.

"At some point it became like a cat-and-mouse game. I would stop at a red light and go. I knew they were there, and they knew I knew they were there. I was concerned about my kids because I didn't know if it was someone that they hired just out of prison that would hurt me or my family," Recarey said.

Despite relentless political pressure, Reiter and Recarey soldiered on, and their determination yielded evidence that supported most of the girls' allegations, the former cops said. They had phone records that showed Epstein and his assistant, Kellen, had called many of the girls. Epstein's flight logs showed that the calls were made when Epstein was in Palm Beach.

They obtained dozens of message pads from his home that read like a who's who of famous people, including magician David Copperfield and Donald Trump, an indication of Epstein's vast circle of influential friends. There were also messages from girls, and their phone numbers matched those of many of the girls Recarey had interviewed, Recarey said. They read: "Courtney called, she can come at 4," or "Tanya can't come at 7 p.m. tomorrow because she has soccer practice."

They also found naked photographs of underage girls in Epstein's closet, Recarey said.

There were also witnesses: Two of Epstein's butlers gave Recarey sworn interviews, confirming that young girls had been coming and going at the house. One of the butlers, Alfredo Rodriguez, told Recarey that when he was tasked with cleaning up the master bath after Epstein's sessions with the girls, he often discovered sex toys. Once, he accidentally stumbled on a high school girl, whom he identified, sleeping naked in Epstein's spa, he testified in a 2009 court deposition.

Rodriguez said he was given the job of paying the girls, telling Recarey that he was "a human ATM machine" because he was ordered by Epstein to keep $2,000 on him at all times. He was also assigned to buy the girls gifts. Rodriguez gave Recarey copies of pages from a book that Epstein and his staff kept with the names and phone numbers for many of the Palm Beach County girls, Recarey said.

Rodriguez, however, held onto the bulk of Epstein's "little black book," and in November 2009 tried to sell it for $50,000 to an undercover FBI agent posing as a victim's lawyer. He was arrested and sentenced in 2012 to federal prison, and died three years later following an illness. The book — listing personal phone numbers for a cavalcade of Epstein's powerful friends and celebrities — eventually became public as part of a civil lawsuit. It listed more than 100 female names and phone numbers under the headings "massage" in every city where Epstein had homes.

In May 2006, Recarey drew up probable cause affidavits, charging Epstein, two of his assistants and one recruiter with sex-related crimes. Instead, Krischer took what Recarey said was the unusual step of referring the case to a state grand jury. Epstein was indicted in state court on a minor charge of solicitation of prostitution.

Recarey said Krischer told him he didn't believe Epstein's accusers, and only two of them were called before the state grand jury investigating the case — even though police had lined up more than a dozen girls and witnesses at that time.

Believing that the case had been tainted, Reiter — that same month, May 2006 — took a very public stance against Krischer, writing a letter, which was released to the news media, calling on Krischer to remove himself from the case. The chief then referred it to the FBI, which opened its own investigation in July 2006, FBI records show.

Reiter said he was effectively blackballed in some Palm Beach circles as a result of going over Krischer's head, and their relationship, once strong, would never be the same.

Reiter has no regrets about what he did.

"There are challenges here that don't exist in a lot of other places because of the affluence in the community, but the only way I could approach this case was that none of that matters. The truth is still the truth. The facts are the facts. Everybody is treated the same."

In the years that followed, several of the victims obtained lawyers and filed civil lawsuits against Epstein. About two dozen lawsuits were filed, starting in 2008. The early cases were particularly brutal for his victims, the court records show. The girls faced fierce grilling from another pack of Epstein's civil attorneys, who questioned them about their boyfriends, drinking, drug use, social media posts, their parents and even their medical histories.

One girl was asked about her abortions, and her parents, who were Catholic and knew nothing about the abortions, were also deposed and questioned.

Licata said the questions from Epstein's civil lawyers were so intimate that she became paranoid that people were following her.

"His lawyers were just in my life inside and out. They asked if I had a baby, if I had an abortion, 'did you sleep with 30 different guys' and 'do you think that played a part?' I said, 'you're going to come at me like that when you represent a guy who is doing this to hundreds of girls? How do you sleep at night?' "

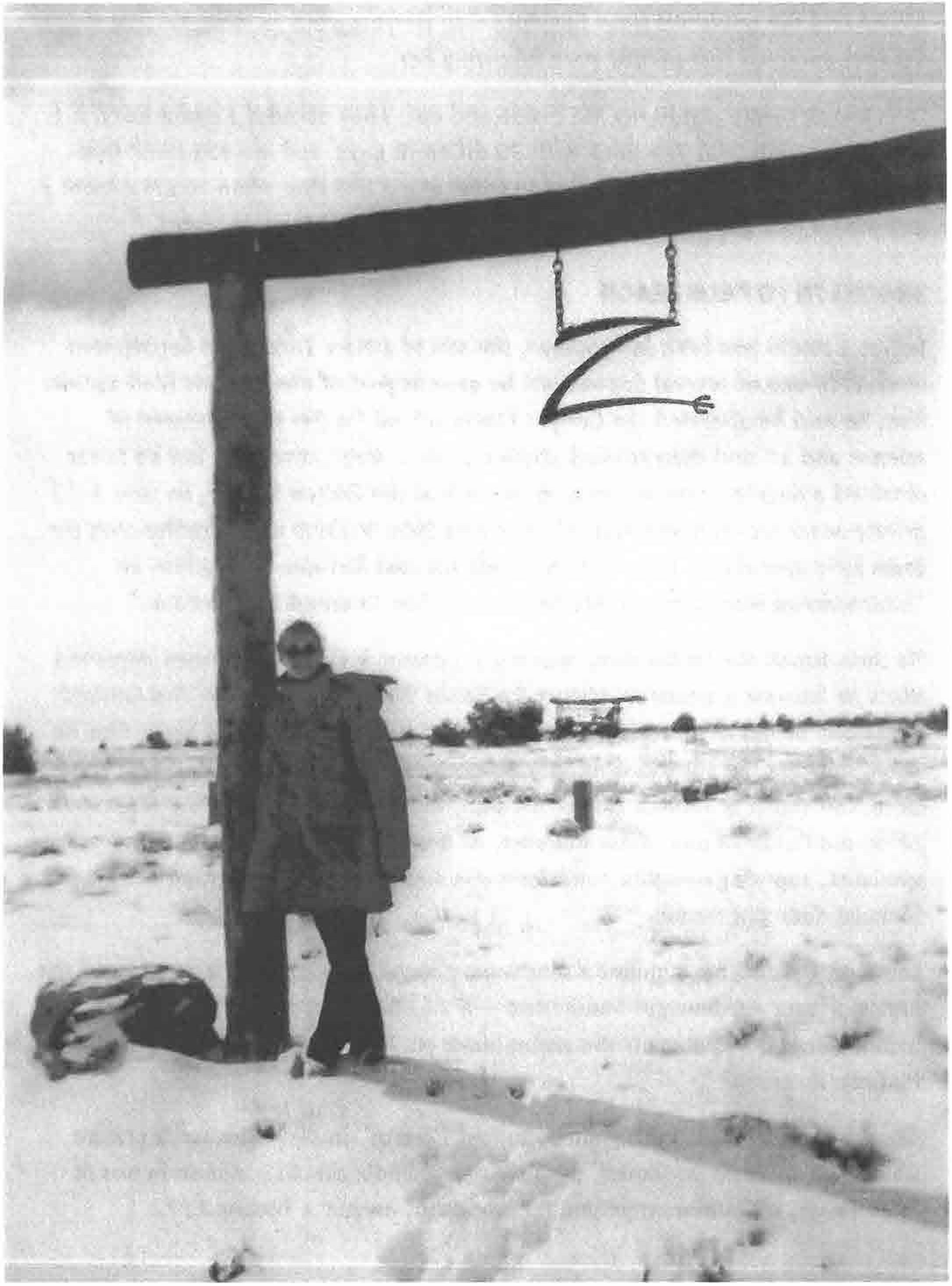
BROOKLYN TO PALM BEACH

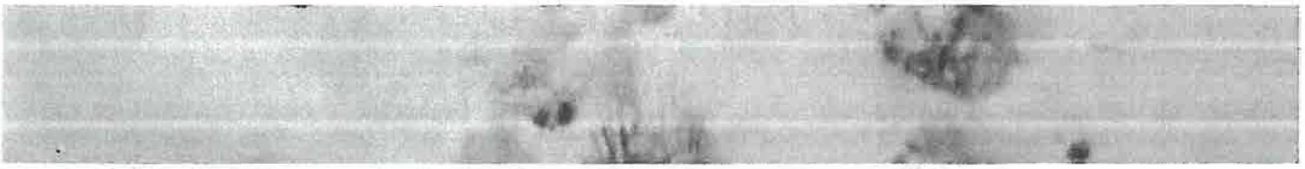
Jeffrey Epstein was born in Brooklyn, the son of a New York parks department worker. In one of several depositions he gave as part of the lawsuits filed against him, he said he attended the Cooper Union school for the advancement of science and art and then studied physics at New York University. But he never obtained a degree, instead going on to teach at the Dalton School, an elite K-12 private academy on Manhattan's Upper East Side. Various news profiles over the years have speculated about how he made his vast fortune, calling him an "International Moneyman of Mystery" and "The Talented Mr. Epstein."

He then struck out on his own, opening J. Epstein & Co. His fortunes improved when he became a financial advisor for Leslie Wexner, founder of The Limited stores and owner of Victoria's Secret brands. Later, Epstein would boast that he would manage the portfolios of only those clients who had $1 billion or more.This much is known: He got his start on Wall Street after being offered a job by the father of one of his students. At Bear Stearns, he became a derivative specialist, applying complex math formulas and computer algorithms to evaluate financial data and trends.

Through Wexner, he acquired a seven-story stone mansion that is considered the largest private residence in Manhattan — a 21,000-square-foot fortress with heated sidewalks that spans the entire block on 71st Street between Fifth and Madison Avenues.

He also owns a 10,000-acre ranch, named "Zorro," in New Mexico, a private island called "Little St. James" in the Virgin Islands, the $13 million house in Palm Beach, a Gulfstream jet and, at one point, owned a Boeing 727.





Virginia Roberts says she was used as a sex slave for Jeffrey Epstein for years starting at the age of 16. Roberts says that Epstein also lent her out to some of his wealthy, powerful acquaintances for sex.
COURTESY OF VIRGINIA ROBERTS

He has never been in the Forbes 400 list of the wealthiest Americans, largely because the magazine has never been able to determine the source or the size of his wealth.

He has been dogged by questions about his financial dealings. A former business partner, Steven Hoffenberg, sued him in 2016, claiming that Epstein was the mastermind behind a $500 million Ponzi scheme that Hoffenberg was imprisoned for in 1995. Hoffenberg served 18 years for the scam, but he later dropped the lawsuit against Epstein.

In August, two of Hoffenberg's former investors rekindled the lawsuit against Epstein, but the case was dropped in October.

Epstein and his associate, British-born socialite Ghislaine Maxwell, were also accused in a 2015 federal civil suit of organizing underage sex parties on his private plane, nicknamed "The Lolita Express," and at Epstein's various homes. Maxwell, who has never been charged with wrongdoing, has denied allegations made in the lawsuit that she was Epstein's "madam." The suit, filed by victim Virginia Roberts, was settled in 2017.

Virginia Roberts was working at Mar-a-Lago when she was recruited to be a masseuse to Palm Beach hedge fund manager Jeffrey Epstein. She was lured into a life of depravity and sexual abuse.

BY EMILY MICHOT ✉ | JULIE K. BROWN ✉

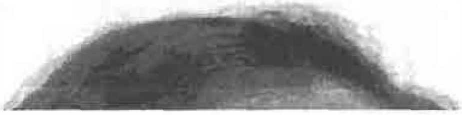
It was Epstein's contacts with powerful and famous people that first propelled him into the public spotlight. In 2002, he flew former President Bill Clinton, actor Kevin Spacey, comedian Chris Tucker and others to South Africa on his private jet as part of a fact-finding AIDS mission in support of the Clinton Foundation.

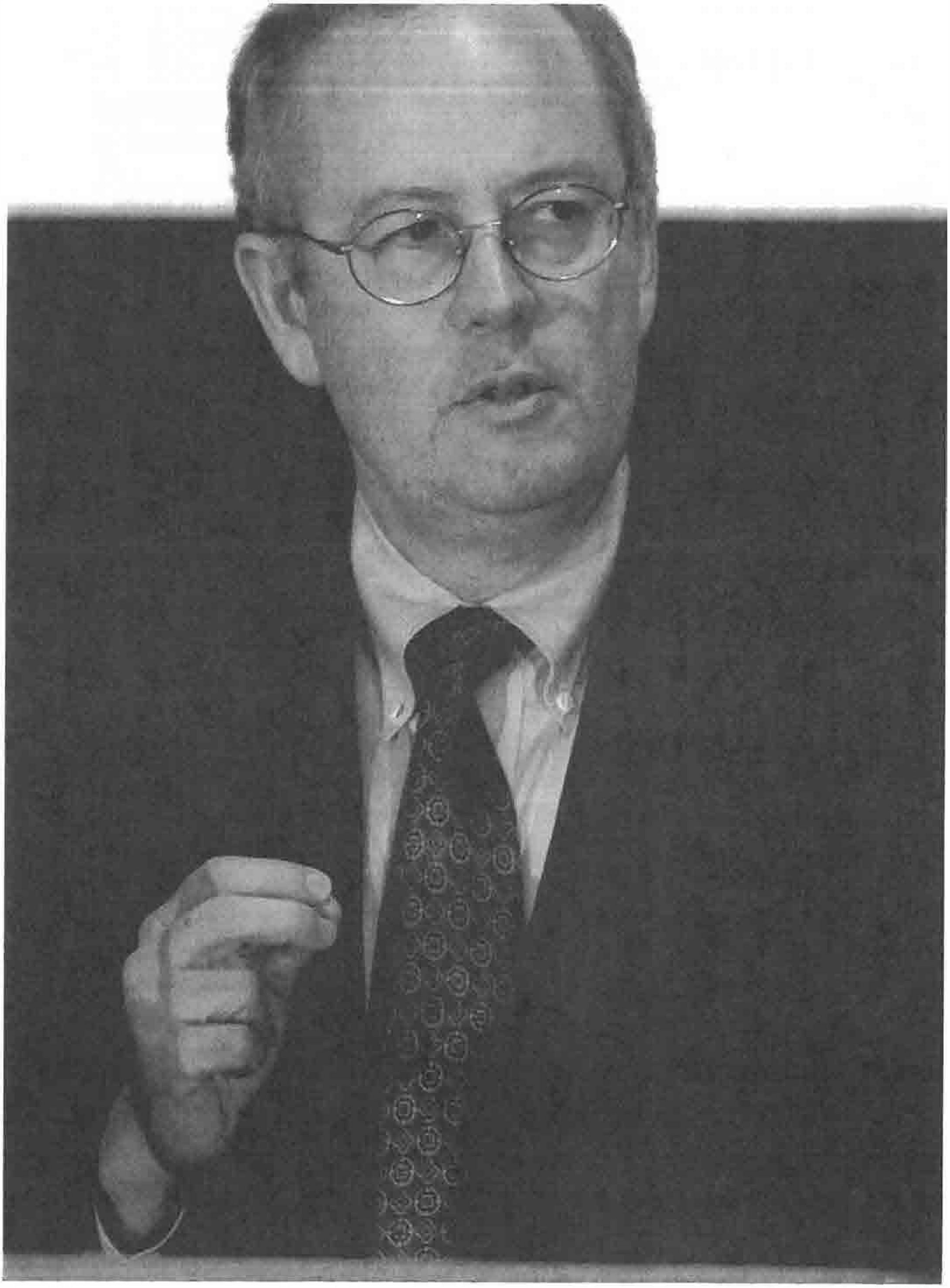
But Epstein, a Clinton donor who contributed hundreds of thousands of dollars to Democratic candidates and causes, realized that his Democratic connections weren't going to help him in 2006, when the federal prosecutor was Acosta, a conservative Republican appointed during the George W. Bush administration.

ENTER KENNETH STARR

Epstein's tactic: hire the most aggressive and politically connected lawyers that his money could buy.

At the top of his list: Kenneth Starr, a Republican icon because of his pursuit of Bill Clinton during the Whitewater investigation, which led to the impeachment (but not conviction) of the president after it was revealed he'd had sex with a young White House intern. Like Acosta, Starr had worked at the prestigious law firm Kirkland & Ellis. Epstein also tapped Jay Lefkowitz, also of Kirkland, who worked as a domestic policy advisor and later as a special envoy to North Korea during the George W. Bush presidency.





Independent Counsel Kenneth Starr speaks to the San Antonio Bar Association in San Antonio, Friday afternoon, May 1, 1998. With his own fight over executive privilege raging in secret, Starr today drew parallels between his plight and Watergate prosecutors. (AP Photo/Eric Gay)
ERIC GAY *AP*

Epstein also hired Bruce Reinhart, then an assistant U.S. attorney in South Florida, now a U.S. magistrate. He left the U.S. Attorney's Office on Jan. 1, 2008, and went to work representing Epstein's employees on Jan. 2, 2008, court records show. In 2011, Reinhart was named in the Crime Victims' Rights Act lawsuit, which accused him of violating Justice Department policies by switching sides, implying that he leveraged inside information about Epstein's investigation to curry favor with Epstein.

Reinhart, in a sworn declaration attached to the CVRA case, denied the allegation, saying he did not participate in Epstein's criminal case and "never learned any confidential, non-public information about the Epstein matter."

The U.S. Attorney's Office has since disputed that, saying in court papers that he did possess confidential information about the case.

Contacted for this story, Reinhart, in an email, said he never represented Epstein — only Epstein's pilots; his scheduler, Sarah Kellen; and Nadia Marcinkova, described by some victims as Epstein's sex slave. Reinhart also pointed out that a complaint filed against him by victims' lawyer Paul Cassell was dismissed by the Justice Department.

Bruce E. Reinhart, Member, McDonald Hopkins LLC. (PRNewsFoto/McDonald Hopkins LLC)

PR NEWSWIRE

That same year, 2011, more girls continued to come forward, including Roberts, who claimed in a British tabloid story that Epstein directed her — while she was underage by Florida standards — to have sex, not only with him, but with other powerful men, including his attorney, Alan Dershowitz, and Prince Andrew. Dershowitz and Andrew denied her claims, but after she filed a sworn affidavit in federal court in Miami, the ensuing news media firestorm forced Acosta, then dean of the law school at Florida International University, to explain why he'd declined to prosecute Epstein.

In a written, public statement on March 20, 2011, Acosta asserted that the deal he struck with Epstein's lawyers was harsher than it would have been had the case remained with the state prosecutor, Krischer, who favored charging Epstein with only a misdemeanor prostitution violation.

Acosta also described what he called a "year-long assault" on prosecutors by Epstein's "army of legal superstars" who, he said, investigated individual prosecutors and their families, looking for "personal peccadilloes" to disqualify them from Epstein's case.

Dershowitz, in an interview, denied that Epstein's lawyers would ever investigate prosecutors.

Documents nevertheless show that Acosta not only buckled under pressure from Epstein's lawyers, but he and other prosecutors worked with them to contain the case, even as the FBI was uncovering evidence of victims and witnesses in other states, FBI and federal court documents show.

A 53-page federal indictment had been prepared in 2007, and subpoenas were served on several of Epstein's employees, compelling them to testify before a federal grand jury. The court records reveal that emails began to fly back and forth between prosecutors and Epstein's legal team. Those emails show that federal prosecutors kept acquiescing to Epstein's demands.

Prosecutors allowed Epstein's lawyers to dictate the terms of each deal that they drew up, and repeatedly backed down on deadlines, so that the defense essentially controlled the pace of the negotiations, the emails and letters show.

It's clear, from emails and other records, that prosecutors spent a lot of time figuring out a way to settle the case with the least amount of scandal. Instead of charging Epstein with a sex offense, prosecutors considered witness tampering and obstruction charges, and misdemeanors that would allow Epstein to secretly plead guilty in Miami instead of in Palm Beach County, where most of the victims lived, thereby limiting media exposure and making it less likely for victims to appear at the sentencing.

"I've been spending some quality time with Title 18 [the U.S. criminal code] looking for misdemeanors," the lead prosecutor, A. Marie Villafaña, wrote to Epstein's lawyers on Sept. 13, 2007, adding that she was trying to find "a factual basis" for one or more non-sex-related crimes to charge him with.

The email chain shows that prosecutors sometimes communicated with the defense team using private emails, and that their correspondence referenced discussions that they wanted to have by phone or in person, so that there would be no paper trail.

"It's highly unusual and raises suspicions of something unethical happening when you see emails that say 'call me, I don't want to put this in writing.' There's no reason to worry about putting something in writing if there's nothing improper or unethical in the case," said former federal prosecutor Francey Hakes, who worked in the Justice Department's crimes against children unit.

On Sept. 24, 2007, another agreement was reached, but Epstein still wasn't happy with it, emails show.

Lefkowitz continued to pressure the U.S. Attorney's Office to keep the agreement secret, even though under the Crime Victims' Rights Act, prosecutors were required to inform the victims that a plea deal had been signed.

"We ... object to your sending a letter to the alleged victims," Lefkowitz wrote on Nov. 28. "... Any such letter would immediately be leaked to the press, your actions will only have the effect of injuring Mr. Epstein and promoting spurious civil litigation directed at him. We also request that if your office believes that it must send a letter to go to the alleged victims ... it should happen only after Mr. Epstein has entered his plea."

The girls who were abused by Jeffrey Epstein and the cops who championed their cause remain angry over what they regard as a gross injustice, while Epstein's employees and those who engineered his non-prosecution agreement have prospered.

BY MARTA OLIVER CRAVIOTTO ✉ | EMILY MICHOT ✉ | JULIE K. BROWN ✉

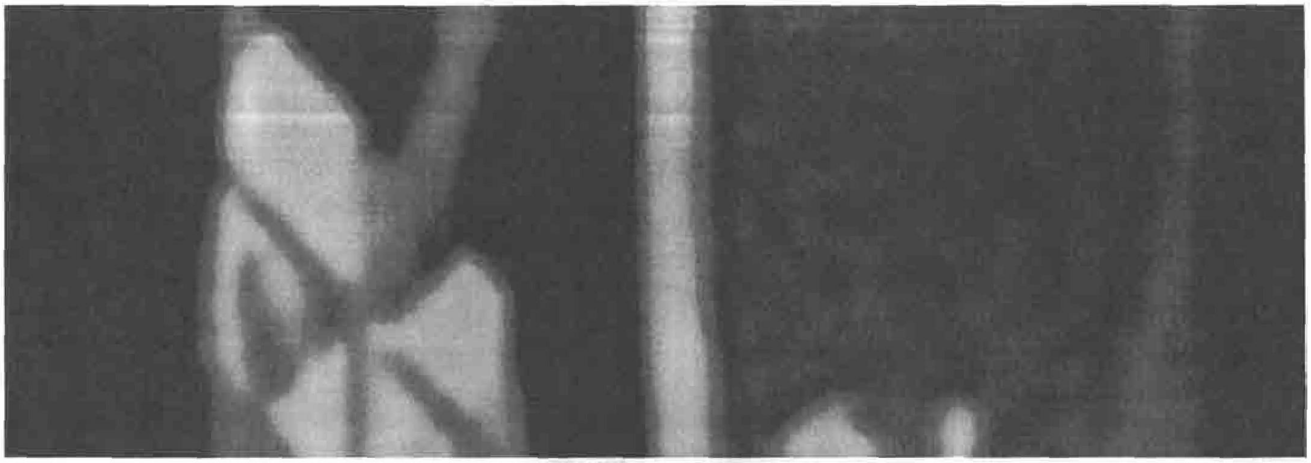
By December, Epstein had still not agreed to a date for his plea hearing, and was technically in violation of the September agreement, which required him to appear in court by November, Acosta noted in a letter to Kenneth Starr in December 2007.

"The [U.S. attorneys] who have been negotiating with defense counsel have for some time complained to me regarding the tactics used by the defense team," Acosta wrote. "It appears to them that as soon as resolution is reached on one issue, defense counsel finds ways to challenge the resolution collaterally. ... Some in our office are deeply concerned that defense counsel will continue to mount collateral challenges to provisions to the agreement, even after Mr. Epstein has entered his guilty plea and thus rendered the agreement difficult, if not impossible, to unwind."

And that's exactly what happened.

Villafaña frequently showed her frustration.

"I thought we had worked very well together in resolving this dispute. ... I feel that I bent over backwards to keep in mind the effect that the agreement would have on Mr. Epstein," Villafaña wrote to Epstein attorney Lefkowitz on Dec. 13, 2007.



By then the deal had been signed for two months, and Jeffrey Sloman, Acosta's top assistant, told Lefkowitz he intended to begin notifying Epstein's victims.

An indignant Lefkowitz wrote to Acosta: "You ... assured me that your office would not ... contact any of the identified individuals, potential witnesses or potential civil claimants and their respective counsel in this matter."

As the months went on, with the agreement still in limbo, federal prosecutors once again began to prepare indictments against Epstein, court records show. The FBI investigation briefly resumed, and additional witnesses were interviewed in New York and New Mexico, the records show. In January 2008, several Epstein victims were sent letters informing them that the FBI investigation was "ongoing" as negotiations to finalize the plea bargain continued behind the scenes.

Starr finally appealed to the Justice Department in Washington, challenging federal jurisdiction of the case, but in May 2008, the Justice Department affirmed Acosta's right to prosecute.

'STILL AFRAID OF EPSTEIN'

In recent court filings, the government was forced to answer questions about its negotiations, finally admitting in 2013 that federal prosecutors had backed down under relentless pressure by Epstein's attorneys.

"The government admits that, at least in part as a result of objections lodged by Epstein's lawyers to victim notifications, the [United States Attorney's Office] reevaluated its obligations to provide notification to victims and Jane Doe #1 was thus not told that the USAO had entered into a non-prosecution agreement with Epstein until after it was signed," wrote Assistant U.S. Attorney Dexter Lee.

Said Hakes, the former federal prosecutor: "I have never heard of a case where federal prosecutors consult with a defense attorney before they send out standard victim notification letters. To negotiate what the letters would say and whether they would be sent at all suggest that the victims' rights were violated multiple times."

Starr's aggressive advocacy for Epstein against allegations of improper sexual behavior was in stark contrast to the path he took investigating then-President Clinton. The Starr Report, the summary of his findings in the Whitewater investigation, which started as a probe of a land deal gone sour and veered into an investigation of sexual misconduct, savaged the president for his involvement with White House intern Monica Lewinsky and was the basis for impeachment.

Starr himself would face criticism in 2016 — he stepped down as president of Baylor University amid allegations that he and other university officials mishandled sexual assault allegations brought by female students against members of the school's football team.

The Herald reached out to Starr, through certified letter and through a spokesman for his current law firm, the Lanier Firm, but did not receive a response for this story.

Palm Beach police detective Recarey, one of the most highly decorated officers on the Palm Beach Police Department, called the Epstein case the most troubling of his 23-year career.

"Some of the victims were — and still are — afraid of Epstein," he said as part of a series of interviews with the Herald earlier this year.

Privately, Reiter and Recarey said, they held onto a hope that Epstein would be brought to trial someday, but they said that that notion had faded.

"I always hoped that the plea would be thrown out and that these teenage girls, who were labeled as prostitutes by prosecutors, would get to finally shed that label and see him go to prison where he belongs," Recarey said.

Recarey died in May after a brief illness. He was 50 years old.

More from the series

PART ONE

How a future Trump Cabinet member gave a serial sex abuser the deal of a lifetime

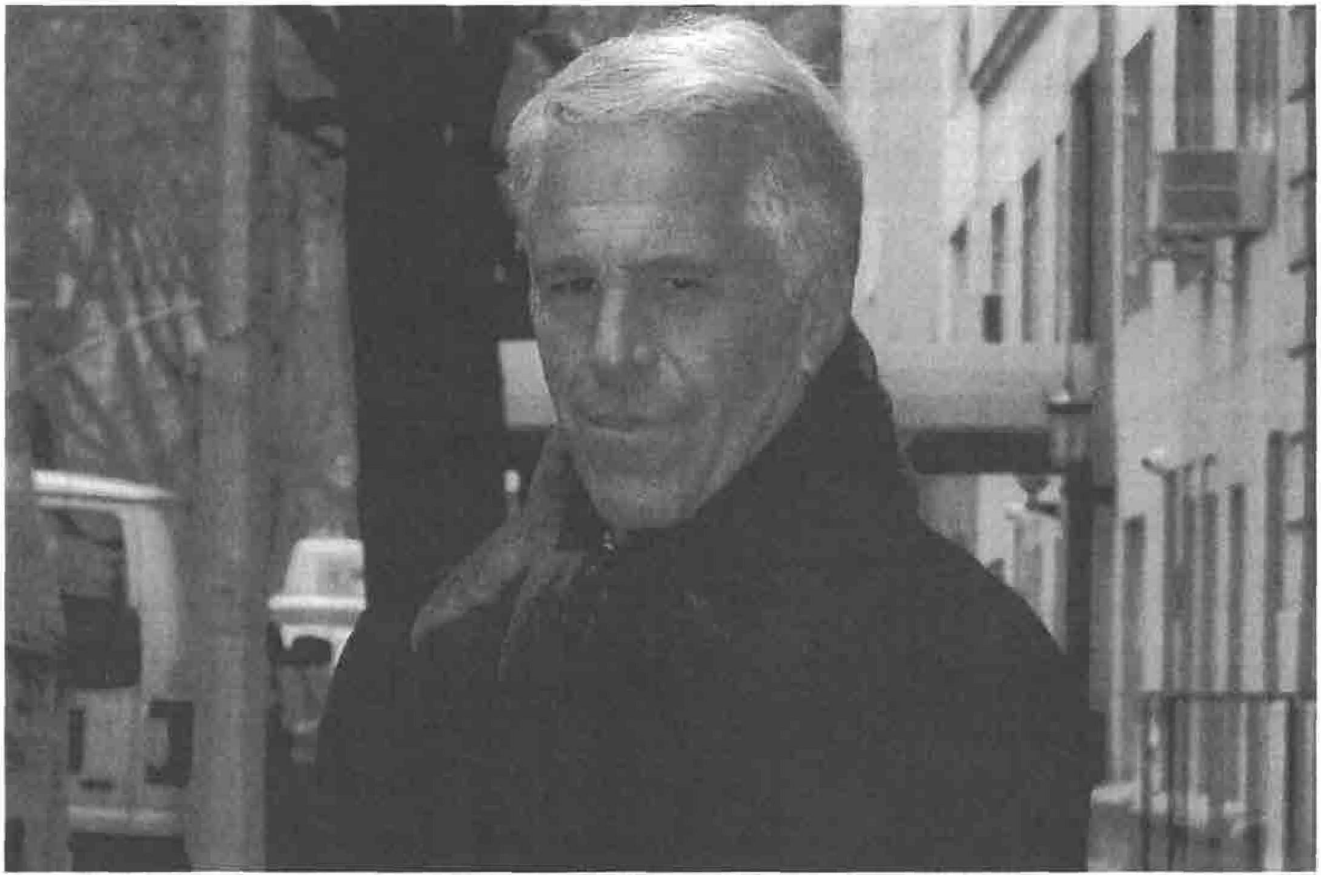
PART TWO

EXHIBIT B

DA's office 'went easy' on sex offender Epstein

By Rebecca Rosenberg and Danika Fears

January 7, 2015 | 2:38am



Jeffrey Epstein
Gregory P. Mango

Jeffrey Epstein may be a convicted pedophile and accused sex- slave master — but that didn't stop him from getting some tender loving treatment from the Manhattan DA's Office.

Prosecutors went to bat for the billionaire pervert at a 2011 legal hearing, asking a judge to cut the filthy- rich felon a break on the severity of his sex- offender status, according to court documents.

ADA Jennifer Gaffney supported a request by Epstein — who allegedly peddled $15,000-a-night teen "sex slaves" to rich men like Britain's Prince Andrew — that he only be listed as a Level 1 pervert during a hearing in Manhattan Supreme Court, according to legal transcripts.

The sicko investment magnate has just finished serving 13 months of an 18-month sentence in Florida for soliciting a minor girl for prostitution. The hearing was to determine what kind of offender status he would have in New York, where he keeps a vacation home.

The city ADA argued that since "there was only an indictment for one victim," Epstein shouldn't have to register as Level 3 — which would require he travel to New York every 90 days and check in with cops.

Judge Ruth Pickholz was flabbergasted. She couldn't understand why the DA would want to go easy on Epstein, after the New York State Board of Examiners of Sex Offenders recommended the Level 3 status because there was more than just the one victim for which he pleaded guilty.

"I have to tell you, I'm a little overwhelmed because I have never seen the prosecutor's office do anything like this," the judge said at the hearing.

Despite the DA's arguments, Pickholz slapped Epstein with Level 3 status, which was upheld on appeal.

"The strong evidence that the offenses against the other victims did occur outweighs any inferences to be drawn from the manner in which this case was prosecuted in Florida," the New York Court of Appeals said in its ruling.

When asked for comment, the DA's Office referred The Post to court documents stating that the prosecution's position was based "largely on the mistaken notion" that only the formal charges against Epstein could be factors in the decision.

The jet-setting financier has been making headlines in recent days after Virginia Roberts, 30, claimed he pimped her to Britain's Prince Andrew.

Prince Andrew has denied the accusations against him.

FILED UNDER **JEFFREY EPSTEIN, PRINCE ANDREW, VIRGINIA ROBERTS**

Recommended by

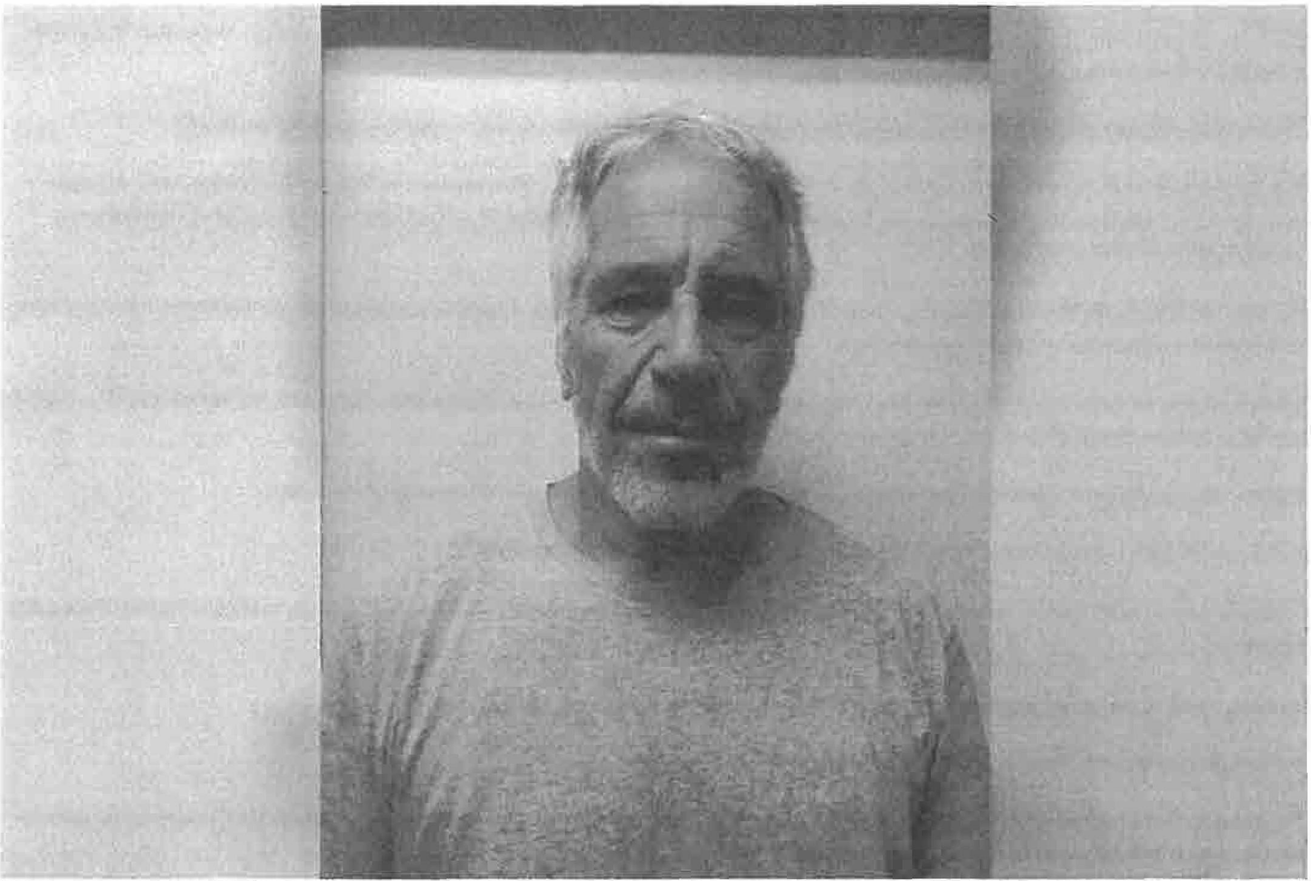
EXHIBIT C

METRO

Manhattan DA sided with pedophile billionaire after botching investigation

By Rebecca Rosenberg, Larry Celona, Susan Edelman and Isabel Vincent

December 1, 2018 | 8:47pm | Updated



Jeffrey Epstein

The Manhattan DA's office once went to bat for billionaire pervert Jeffrey Epstein, after botching a review of his sex crimes and swallowing his lawyers' claim that "there are no real victims here," records obtained by the Post show.

Assistant DA Jennifer Gaffney, then-deputy chief of Cyrus Vance Jr.'s sex-crimes unit, in January 2011 asked a Manhattan judge to downgrade Epstein's status in the New York sex-offender registry from the most-dangerous Level 3 to least-restrictive Level 1.

The judge was stunned.

"I have never seen the prosecutor's office do anything like this," Manhattan Supreme Court Justice Ruth Pickholz told Gaffney. "I have done many [cases] much less troubling than this one where [prosecutors] would never make a downward argument like this."

Pressed by the judge, Gaffney admitted that she never spoke to the Florida U.S. Attorney who handled a sprawling sex-crime investigation into the financier.

"I don't think you did much of an investigation here," Pickholz said. "I am shocked."

Vance's mishandling of the Epstein hearing has come under new scrutiny after a Miami Herald report last week revealed a secret "non-prosecution agreement" in Florida that buried evidence Epstein had allegedly pimped out 80 girls and young women to his rich and powerful pals.

The DA's office insists Vance "was not aware" of the hearing until years later and had nothing to do with it.

"Our prosecutor made a mistake," Vance spokesman Danny Frost said of Gaffney.

Gaffney, a prosecutor working in the Harvey Weinstein probe, left the DA's office in September. She declined to comment.

A DA insider said the office was unaware of Epstein's secret plea deal in Florida, and never investigated his sexcapades in NYC.

Some law enforcement sources don't believe Vance had no clue that his office had a sex-offender case involving a Manhattan mogul with close ties to Democrats.

"This is very unusual," one said. "There is no way Vance didn't know. The question is why — and who asked for the favor."

The FBI found Epstein recruited girls as young as 13, many runaways, from Florida, New Mexico, the Caribbean and New York, and paid them for nude massages that often led to sex, the Herald reported. Some girls worked for a Manhattan modeling agency and lived in a nearby apartment owned by Epstein.

But Miami U.S. Attorney Alexander Acosta — now President Trump's labor secretary — sealed a 53-page federal indictment that could have sent Epstein to prison for life, the Herald reported.

Instead, Acosta let Epstein plead guilty to procuring a person under 18 for prostitution. Sentenced to 18 months, he served just 13 — most of it in his tony Palm Beach office.

Epstein was required to register as a sex offender in New York because one of his many homes is in Manhattan.

In the Jan. 18, 2011 hearing, Gaffney argued the evidence didn't justify the harshest status.

"There is only an indictment for one victim,," she said. "If an offender is not indicted for an offense, it is strong evidence that the offense did not occur."

Pickholz rejected Gaffney's arguments and gave Epstein the highest sex-offender status – Level 3.

When Epstein appealed, Vance's office admitted it had misread the law.

The appellate court upheld Pickholz, saying Epstein "committed multiple offenses against a series of underage girls," and that the victim in his indictment "was only one of defendant's many victims."

FILED UNDER **BILLIONAIRES, DISTRICT ATTORNEYS, JEFFREY EPSTEIN, PEDOPHILES, SEX OFFENDERS**

Recommended by

EXHIBIT D

From: **Frost, Danny** <frostd@dany.nyc.gov>
Date: Thu, Dec 6, 2018 at 5:27 PM
Subject: RE: request
To: Susan Edelman <sedelman@nypost.com>

Hi Sue,

Regarding your first question: pursuant to Civil Rights Law § 50-b, our office's practice in appellate sex crimes matters is to file documents under seal. This is not particular to People v. Epstein; it is routine across our appellate filings. My understanding is that it is also the Appellate Division's (the court's) practice to maintain such seal.

That being said, I have discussed your second question with our attorneys. If the Post petitions the court, and the court asks the People for our positon, we will *not* oppose the petition for a redacted brief.

Thanks.

Danny Frost
Director of Communications
Manhattan District Attorney Cy Vance, Jr.
212-335-9400 // @ManhattanDA

From: Susan Edelman [mailto:sedelman@nypost.com]
Sent: Thursday, December 6, 2018 9:12 AM
To: Frost, Danny <frostd@dany.nyc.gov>
Subject: Re: request

Hi Dan,

Please explain why the DA's office didn't file a redacted brief, but rather covered up all the facts.

If we filed a petition for a redacted brief, would the DA's office fight it or cooperate?

Thank you,

Sue

Susan Edelman

Reporter

1211 Avenue of the Americas, New York, NY, 10036

917-584-7675

sedelman@nypost.com

On Wed, Dec 5, 2018 at 3:45 PM Frost, Danny <frostd@dany.nyc.gov> wrote:

Hi Sue – the document was filed under seal. I cannot violate the seal and provide it to you.

(off the record please) You may wish to petition the court, or have the Post's lawyer send our lawyers something. We've reached the end of me being able to help – it's sealed.

Danny Frost
Director of Communications
Manhattan District Attorney Cy Vance, Jr.
212-335-9400 // @ManhattanDA

From: Susan Edelman [mailto:sedelman@nypost.com]
Sent: Wednesday, December 5, 2018 3:06 PM
To: Frost, Danny <frostd@dany.nyc.gov>
Subject: Re: request

Right, but the law allows for the redaction of victim names. Where does it say the whole document is denied?

Susan Edelman

Reporter

1211 Avenue of the Americas, New York, NY, 10036

917-584-7675

sedelman@nypost.com

On Wed, Dec 5, 2018 at 3:02 PM Frost, Danny <frostd@dany.nyc.gov> wrote:

Hi Sue - this is prescribed by Civil Rights Law § 50-b.

Danny Frost
Director of Communications
Manhattan District Attorney Cy Vance, Jr.
212-335-9400 // @ManhattanDA

From: Susan Edelman [mailto:sedelman@nypost.com]
Sent: Wednesday, December 5, 2018 2:58 PM
To: Frost, Danny <frostd@dany.nyc.gov>
Subject: Re: request

Hi Dan,

So it was a unilateral decision by the DA to seal the document?

There was no hearing or opportunity to object?

Thank you,

Sue

Susan Edelman

Reporter

1211 Avenue of the Americas, New York, NY, 10036

917-584-7675

sedelman@nypost.com

On Wed, Dec 5, 2018 at 2:55 PM Frost, Danny <frostd@dany.nyc.gov> wrote:

Hi Sue – we do not have a court order; rather, we are bound by the Civil Rights statute cited on the cover page. Thanks.

Danny Frost
Director of Communications
Manhattan District Attorney Cy Vance, Jr.
212-335-9400 // @ManhattanDA

From: Susan Edelman [mailto:sedelman@nypost.com]
Sent: Wednesday, December 5, 2018 1:47 PM
To: Frost, Danny <frostd@dany.nyc.gov>
Subject: Re: request

Thanks,, Is there a specific date?

Do you have a court order sealing the document?

Susan Edelman

Reporter

1211 Avenue of the Americas, New York, NY, 10036

917-584-7675

sedelman@nypost.com

On Wed, Dec 5, 2018 at 1:44 PM Frost, Danny <frostd@dany.nyc.gov> wrote:

August 2011

Danny Frost
Director of Communications
Manhattan District Attorney Cy Vance, Jr.
212-335-9400 // @ManhattanDA

From: Susan Edelman [mailto:sedelman@nypost.com]
Sent: Wednesday, December 5, 2018 1:38 PM
To: Frost, Danny <frostd@dany.nyc.gov>
Subject: Re: request

P.S. What is the date of this document?

Susan Edelman

Reporter

1211 Avenue of the Americas, New York, NY, 10036

917-584-7675

sedelman@nypost.com

On Wed, Dec 5, 2018 at 1:02 PM Susan Edelman <sedelman@nypost.com> wrote:

Thanks, Danny,

Is there a judge's order sealing this document?

Or is it just stated by the DA's office.

It's unclear on this cover page.

Sue

Susan Edelman

Reporter

1211 Avenue of the Americas, New York, NY, 10036

917-584-7675

sedelman@nypost.com

On Wed, Dec 5, 2018 at 12:55 PM Frost, Danny <frostd@dany.nyc.gov> wrote:

Attached, thanks.

Danny Frost
Director of Communications
Manhattan District Attorney Cy Vance, Jr.
212-335-9400 // @ManhattanDA

From: Susan Edelman [mailto:sedelman@nypost.com]
Sent: Wednesday, December 5, 2018 12:41 PM
To: Frost, Danny <frostd@dany.nyc.gov>
Subject: Re: request

Hi Dan,

It's for my information -- not publication.

Thanks much,

Sue

Susan Edelman

Reporter

1211 Avenue of the Americas, New York, NY, 10036

917-584-7675

sedelman@nypost.com

On Wed, Dec 5, 2018 at 12:38 PM Frost, Danny <frostd@dany.nyc.gov> wrote:

Hi Sue – would I be sending the cover page for publication, or for your information? I am happy to send it to you not-for-publication, but if you intend to publish it, I need to check with our lawyers. Thanks.

Danny Frost
Director of Communications
Manhattan District Attorney Cy Vance, Jr.
212-335-9400 // @ManhattanDA

From: Susan Edelman [mailto:sedelman@nypost.com]
Sent: Wednesday, December 5, 2018 12:35 PM
To: Frost, Danny <frostd@dany.nyc.gov>
Subject: Re: request

Hi Danny,

Thanks. Can you send me a PDF of the cover page?

Who issued that order?

We always get such court documents -- with the names and ID of victims redacted.

The whole document is not withheld.

Sue

Susan Edelman

Reporter

1211 Avenue of the Americas, New York, NY, 10036

917-584-7675

sedelman@nypost.com

On Wed, Dec 5, 2018 at 12:28 PM Frost, Danny <frostd@dany.nyc.gov> wrote:

Hi Sue – I've confirmed that the brief is filed under seal. It contains the following header on the cover page: "Pursuant to Civil Rights Law § 50-b, the identities of the victims, who are the victims of sex offenses, shall be confidential, **and this document shall not be made available for public inspection."**

Accordingly we must decline your request. Thanks.

Danny Frost
Director of Communications
Manhattan District Attorney Cy Vance, Jr.
212-335-9400 // @ManhattanDA

From: Susan Edelman [mailto:sedelman@nypost.com]
Sent: Tuesday, December 4, 2018 6:34 PM
To: Frost, Danny <frostd@dany.nyc.gov>
Subject: request

Hi Danny,

Following up on the Epstein case:

Can you send me the DA's appellate brief in the sex-offender registry case?
I understand you may have to redact the names of any victims.

Please let me know as soon as possible.

Thank you,

Sue

Susan Edelman

Reporter

1211 Avenue of the Americas, New York, NY, 10036

917-584-7675

sedelman@nypost.com

This email communication and any files transmitted with it contain privileged and confidential information from the New York County District Attorney's Office and are intended solely for the use of the individuals or entity to whom it has been addressed. If you are not the intended recipient, you are hereby notified that any dissemination or copying of this email is strictly prohibited. If you have received this email in error, please delete it and notify the sender by return email.

EXHIBIT E

SUPREME COURT OF THE STATE OF NEW YORK
APPELLATE DIVISION: FIRST DEPARTMENT

THE PEOPLE OF THE STATE OF NEW YORK, Respondent,	AFFIRMATION IN RESPONSE TO MOTION
-against-	
JEFFREY E. EPSTEIN,	N.Y. Co. Ind. No. 30129/10
Defendant-Appellant.	Cal. Date: December 31, 2018

KAREN FRIEDMAN AGNIFILO, an attorney duly admitted to practice before the courts of this State, affirms under penalty of perjury that:

1. I am an Assistant District Attorney, of counsel to CYRUS R. VANCE, JR., District Attorney of New York County, and I submit this affirmation on behalf of the People in response to a motion filed by non-party movant NYP Holdings, Inc. (the "Post") to "unseal" the appellate briefs from 2011, pertaining to the adjudication of defendant Jeffrey E. Epstein as a Level III sex offender under the Sex Offender Registration Act ("SORA"). Civil Rights Law § 50-b provides that the identities of the victims of sex offenses must remain confidential and that documents bearing identifying information should not be made available for public inspection. Despite those protections, the Post seeks copies of the appellate briefs, with the names or initials of Epstein's sex crimes victims redacted, for use in reporting about the procedural history underlying defendant's SORA adjudication.

2. In order to protect the privacy interests of victims of sex crimes, the plain language of Civil Rights Law § 50-b prohibits production to the public of the appellate briefs, which would reveal the identities of many of defendant's sex crime victims. To be sure, the statute provides for a narrow exception to the confidentiality requirement, permitting disclosure, by court order, of information for "good cause" and upon "notice to the victim... and the public officer or employee charged with the duty of prosecuting the offense." Civil Rights Law § 50-b(2)(b). We note that the prosecuting agencies for defendant's underlying sex crimes are the federal and local prosecutor's offices in Florida, where defendant was charged. By all appearances, the Post has yet to furnish the requisite notice to either the appropriate prosecuting agencies or to the victims themselves. And to the extent that the Post is unable to furnish notice to the victims, this Office, which was not the prosecuting agency, is not in a position to do so.

3. In any event, and in keeping with the People's obligation under Civil Rights Law § 50-b to protect the privacy of the victims of sex crimes, we cannot agree to the wholesale production of the People's appellate brief or even to a production of the People's brief with redactions of the names or initials of Epstein's victims.[1] However, if this Court is inclined to grant the Post's motion, we would not oppose producing a copy of the People's brief, with substantial redactions necessary to protect the identities of the victims but keeping

---

[1] According to the Post's filing, they have contacted defendant's current counsel, who has reserved the right to oppose the disclosure of defendant's appellate brief.

intact those portions of the brief that recount the procedural history of the SORA hearing, the portion of the brief that is seemingly of interest to the Post.

Dated: New York, New York
December 28, 2018

KAREN FRIEDMAN AGNIFILO
Assistant District Attorney
(212) 335-9000

cc: Davis Wright Tremaine LLP
John M. Browning
1251 Avenue of the Americas, 21st Floor
New York, New York 10020-1104